UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS HUMANAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

SUMÁRIO

Table of Contents

EDITORIAL:
PERSPECTIVAS EM FILOSOFIA DA MENTE E DA LINGUAGEM

EDITORIAL:
PERSPECTIVES ON THE PHILOSOPHY OF MIND AND LANGUAGE

CÉSAR MEURER[1]
Universidade Federal de Uberlândia (UFU) - Brasil
cesarmeurer@gmail.com

NARA M. FIGUEIREDO[2]
Universidade Federal de São Paulo (Unifesp) - Brasil
naramfigueiredo@gmail.com

RAQUEL KREMPEL[3]
Universidade Federal de São Paulo (Unifesp) - Brasil
raquelak@gmail.com

Os artigos que compõem este dossiê lidam com temas centrais da filosofia da mente e da filosofia da linguagem. Alguns deles abordam temas clássicos e apresentam panoramas em língua portuguesa, de modo que se fazem úteis a jovens leitores. Outros textos apresentam propostas mais ousadas e prometem incitar debate entre filósofos mais experientes. Desse modo, este volume se equilibra ao oferecer conteúdo para diversos públicos. Esta introdução pretende fornecer uma descrição sintética das contribuições deste volume, para que o leitor possa ter uma visão sinóptica do conteúdo apresentado. Em filosofia da mente, há discussões sobre o consciente e o inconsciente, qualia, fisicalismo, alucinações, hipótese da mente estendida, emoções, o problema da lacuna explicativa e reducionismo nas explicações em ciências cognitivas. Em filosofia da linguagem, são tratados temas como a relação entre pensamento e linguagem, normatividade linguística e demonstrativos. Além disso, contamos com textos com abordagens históricas que tratam dos escritos de Wittgenstein acerca da psicologia e da investigação de Antístenes acerca dos nomes. Ao longo dos próximos parágrafos, apresentamos algumas ideias-chave de cada contribuição, seguindo uma ordem temática.

[1] Pesquisador de Pós-doutorado em Filosofia - UFU/CAPES.
[2] Pesquisadora de Pós-doutorado em Filosofia - Unifesp/CAPES.
[3] Pesquisadora de Pós-doutorado em Filosofia - Unifesp/FAPESP, processo 2018/12683-9.

No artigo “A consciência como uma percepção do mental e o estatuto dos fenômenos mentais inconscientes na perspectiva de David Armstrong”, Tárik de Athayde Prata apresenta e discute as visões do filósofo David Armstrong sobre a consciência e sobre o inconsciente. Segundo o autor, Armstrong é um expoente da teoria de ordem superior sobre estados conscientes. De acordo com a teoria de Armstrong, alguns estados mentais possuem a propriedade de serem conscientes quando são monitorados por outro estado mental (de segunda ou terceira ordem), que se dirige ou é sobre o primeiro. Essa concepção de consciência, que Armstrong chama de consciência introspectiva, considera que a consciência é uma forma de percepção de estados mentais internos. Conforme observa Prata, ela é anti-cartesiana, porque considera que não é essencial a um estado mental que ele seja consciente. O autor expõe também outras duas concepções de consciência de Armstrong: consciência mínima e consciência perceptiva. A consciência mínima é o tipo de consciência possuída, por exemplo, por alguém em um sono sem sonhos. Apesar de receber esse nome, o autor observa que nessa condição a criatura está inconsciente, e tem estados mentais inconscientes. Já a consciência perceptiva ocorre quando se tem percepções do ambiente ou do próprio corpo. A consciência introspectiva seria, contudo, o tipo central. Segundo Armstrong, ela integraria estados e atividades mentais, tornando nossa atividade mental mais sofisticada e permitindo ações mais complexas. Prata termina o artigo observando, contudo, que essa visão sugere que estados inconscientes teriam um papel menos importante em nossas vidas mentais. Segundo ele, isso estaria em desacordo com o que revelam pesquisas em psicologia social, por exemplo, que mostram a relevância de estados inconscientes na tomada de decisão.

No texto “How to measure a Quale” Osvaldo Frota Pessoa Jr. discute a possibilidade de se medir uma qualidade fenomênica. Para isso, o autor primeiramente contextualiza o fisicalismo qualitativo, que defende a ideia de que as qualidades de nossas experiências fenomênicas são físicas, e, portanto, mensuráveis. Em seguida, Pessoa esclarece que ser físico significa envolver processos no espaço e no tempo, em escalas mensuráveis. Seu próximo passo é defender a hipótese de que a consciência é estrutural e material. A estrutura é dada pela distribuição dos neurônios e demais células e moléculas, bem como por sua interação, enquanto o material é a característica dos sistemas biológicos que possibilita a consciência. Pessoa considera então possíveis localizações cerebrais do campo visual subjetivo e segue afirmando que medir os padrões de ativação no tecido cerebral correspondente aos qualia não é capturá-los. No entanto, o autor sugere que medir um quale não é o mesmo que medir os aspectos estruturais e organizacionais usuais da física. Isso se deve ao fato de que não podemos capturar qualia em si mesmos, apenas a partir de um meio, a saber, a materialidade dos sistemas biológicos. Apesar disso, as medições feitas a partir do modelo fisicalista qualitativo permitem, em tese, que um quale seja reproduzido em outro meio igualmente apto. Para ilustrar isso, o autor considera um exemplo de qualia invertidos. Ao final, Pessoa reitera a ideia de que o nosso campo visual ocupa uma pequena parte de nosso cérebro, fazendo alusão aos aspectos temporal e espacial da percepção, e conclui que a questão da medida

dos qualia é resolvida de modo direto, por meio do reconhecimento de que meios igualmente aptos geram qualia similares.

Em "Fisicalismo e o problema mente-cérebro: uma questão de definição", Júlio César Martins Mazzoni oferece uma revisão crítica das definições de 'físico' encontradas nas principais filosofias fisicalistas da mente. Segundo Mazzoni, as distintas formas de definir e conceber o 'físico' podem ser agrupadas em quatro categorias: "(1) Definições de apelo à Física presente, também conhecidas por presentistas, que alegam ser físico somente o que a Física atual postula em seu sistema teórico como tal; (2) Definições de apelo à uma Física ideal, conhecida por futuristas, cuja afirmação é de que físico é tudo aquilo contido numa compreensão completa do Universo de uma Física ideal do futuro; (3) Definição negativa, também conhecida como via negativa, a qual diz ser físico tudo aquilo que é não-fundamentalmente mental; e (4) Definições híbridas e alternativas, compostas pelos critérios de definição das opções anteriores combinadas" (Mazzoni, neste volume). Em seguida, Mazzoni faz um balanço bem informado das principais objeções que podem ser levantadas à definição de físico. "Até a presente data", ele observa, "não está claro o quanto e se tais objeções foram adequadamente respondidas ou superadas" (neste volume). Diante desse estado de coisas, Mazzoni conclui que "não existe uma definição de físico amplamente aceita que não enfrente uma série considerável de dificuldades teóricas". Por conta disso, ele aconselha "que se busque uma resposta a essa questão antes de dar continuidade às discussões sobre o valor de verdade das divergentes alternativas filosóficas apresentadas como soluções ao problema mente-cérebro" (Mazzoni, neste volume).

Marco Aurélio Sousa Alves, no texto "The obscure content of hallucination", propõe uma comparação entre as visões de Tye e de Johnston sobre as alucinações, apontando para diversas semelhanças entre elas. Ele inicia caracterizando versões fortes e fracas do conjuntivismo e do disjuntivismo. Para ele, tanto Tye como Johnston são adeptos de um disjuntivismo e de um conjuntivismo fracos. Eles aceitam que a percepção verídica e a alucinação têm conteúdos diferentes, ao mesmo tempo em que reconhecem que há algo de comum entre elas (mas não algo mental). Ambos aceitam também que objetos singulares são constitutivos da percepção verídica e não das alucinações. Nas alucinações, estamos em contato com propriedades, concebidas como universais não instanciados. Assim, eles pretendem preservar uma forma de realismo direto na experiência verídica, em face ao argumento da alucinação, ao mesmo tempo em que tentam explicar a semelhança de caráter fenomênico entre percepção verídica e alucinação. Alves considera também a concepção epistêmica da alucinação, que explora a ideia de que percepção verídica e alucinação são indistinguíveis não porque compartilham o mesmo caráter fenomênico, mas porque o sujeito que alucina não consegue saber por introspecção que a sua experiência não é verídica. Essa abordagem, adotada por disjuntivistas radicais, é criticada pelo autor. Por fim, Alves levanta algumas dificuldades para a ideia de que nas alucinações estamos em contato com universais não instanciados, a qual está na base da tentativa (tanto de Tye como de Johnston) de combinar um

conjuntivismo fraco com o realismo direto. Segundo o autor, somos levados a um dilema: os disjuntivistas radicais não conseguem explicar a indiscriminabilidade subjetiva entre a percepção verídica e a experiência alucinatória, e os conjuntivistas não conseguem explicar o que a percepção verídica e a alucinação têm em comum.

No artigo “The extended mind hypothesis: an anti-metaphysical vaccine”, Giorgio Airoldi trata da hipótese da mente estendida, primeiro formulada por Clark e Chalmers em 1998. Segundo ela, estados cognitivos podem ser constituídos não só por estados cerebrais, como também por elementos externos ao sujeito. Clark e Chalmers ilustram essa ideia com o exemplo de Otto, um indivíduo com Alzheimer que usa um caderno para ajudá-lo a chegar a um museu. O caderno desempenha a mesma função que a memória, e seria um constituinte da mente. Eles adotam o princípio de paridade, segundo o qual é a função, e não a localização, o que caracteriza processos cognitivos. A distinção entre interno e externo deixa de ser importante. Desse modo, elementos do ambiente podem ser constitutivos de sistemas cognitivos, e não apenas causalmente relevantes para eles. A hipótese da mente estendida é uma das manifestações do externalismo e do funcionalismo. Airoldi observa que há também uma segunda corrente que aceita a hipótese da mente estendida, que procura dar conta de algumas dificuldades relacionadas ao princípio de paridade. Seus proponentes adotam o princípio de complementaridade, segundo o qual elementos externos são complementares aos elementos internos, e não isomórficos a eles. O autor destaca também visões diferentes sobre o escopo da hipótese. Originalmente ela foi aplicada a estados intencionais como crenças, mas se discute também se ela seria aplicável a outros estados, como percepções e estados conscientes. Chalmers, por exemplo, defende um internalismo com relação à consciência. Há também visões com diferentes graus de comprometimento com a tese de que a mente se estende para algo externo, indo do internalismo (que nega essa tese) ao externalismo radical da hipótese da mente estendida, passando por algumas visões intermediárias. O autor revisita diversas críticas contra essa hipótese, bem como as respostas de seus proponentes, observando que muitas críticas se apoiam em concepções questionáveis sobre a marca do mental, ou sobre a distinção entre processos internos e externos. Após apresentar a hipótese e muito do debate em torno dela, o autor conclui que a sua principal virtude é a de servir como uma espécie de vacina contra certos preconceitos metafísicos sobre a estrutura da mente, sua função e seus limites.

No artigo "Emoções corporificadas: uma perspectiva sistêmica sobre estados emocionais", Matheus de Mesquita Silveira pretende mostrar a possibilidade de compreender emoções a partir de uma perspectiva sistêmica. Para cumprir seu objetivo, ele se baseia em dois pontos principais da literatura sobre a corporeidade das emoções: (1) a premissa de Lange segundo a qual alterações periféricas no corpo influenciam as emoções (chamada de visão periférica); e (2) a premissa de James, que consiste na ideia de que estados emocionais podem ser reduzidos a percepções corporais (chamada de visão

perceptiva). Para tratar da visão perceptiva, o autor opta por considerar a teoria de Prinz, representativa dessa visão, que sugere que emoções são estados que registram mudanças corporais. Primeiramente, o autor apresenta evidências empíricas em favor da visão periférica, que sugerem que expressões faciais, postura corporal e respiração, por exemplo, interferem nas emoções. A seguir, ele apresenta críticas à visão perceptiva e à visão periférica, e respectivos contra-argumentos. Por fim, o autor caracteriza e defende a visão sistêmica, que não seria vulnerável às críticas que se configuram a partir de uma perspectiva corporificada das visões periférica e perceptiva. Nessa visão, as emoções são constituídas tanto por estados periféricos corporificados quanto por estados cerebrais. A perspectiva sistêmica pode ser considerada uma variação da visão periférica, e é defendida por ser uma forma mais adequada de explicar a ideia de que alterações periféricas influenciam emoções. Silveira conclui que (1) há evidências empíricas em favor da tese de que a manipulação periférica influencia emoções; e (2) que as emoções podem ser consideradas estados que resultam da integração dos sistemas responsáveis pela motivação, preparação da ação e expressão.

No artigo "A critical approach to sensorimotor contingency theory: brain as agent and conscious mind as a guide of action", Jonas Gonçalves Coelho apresenta alguns aspectos da teoria da contingência sensoriomotora, elaborada por O'Regan e Nöe. Essa teoria é uma tentativa de solucionar o problema da lacuna explicativa, que consiste na dificuldade de se explicar como estados e processos físicos podem dar origem a experiências subjetivas. Conforme expõe Coelho, os autores argumentam que esse problema afeta concepções representacionalistas da experiência consciente, porque elas adotam a noção de qualia. A teoria sensoriomotora, por outro lado, segundo seus proponentes, não estaria sujeita a esse problema, porque rejeita os qualia e concebe a experiência consciente em termos das relações do indivíduo com o ambiente externo. Essas relações são mediadas por leis da contingência sensoriomotora. O cérebro teria, para O'Regan e Nöe, um papel secundário no que diz respeito à experiência consciente. No entanto, conforme observa Coelho, uma série de objeções foram levantadas contra essa teoria, em especial contra a sua pretensa solução do problema da lacuna explicativa. Coelho sugere então outra maneira de lidar com esse problema. Assim como O'Regan e Nöe, ele aceita a importância do ambiente externo e do corpo para o surgimento da experiência consciente, mas enfatiza que o cérebro tem um papel mais central. Segundo ele, a consciência seria uma propriedade não física do cérebro. No caso da consciência visual, o cérebro a usaria "como guia para iniciar e manter ações adaptativas no ambiente em que vive" (Coelho, neste volume).

No texto "Neo-mechanistic explanatory integration for cognitive science: the problem of reduction remains", Diego Azevedo Leite analisa e compara duas visões influentes sobre explicações em ciências cognitivas: explicações neomecanicistas e explicações reducionistas. Segundo ele, os desacordos entre elas estão (1) no modo como concebem as relações entre os vários níveis de explicação da cognição e (2) na concepção de qual desses níveis é mais

explicativo. Para os defensores do neomecanicismo, haveria uma pluralidade de níveis de causação e de explicação científica do cérebro e da cognição, envolvendo mecanismos, suas partes e modo de organização. Alguns autores argumentam que as explicações de nível mais alto não seriam redutíveis às explicações de nível mais baixo, já que o comportamento de um mecanismo como um todo não poderia ser reduzido ao de suas partes. Haveria então algum tipo de autonomia dos níveis mais altos de explicação, que seriam os mais relevantes. Leite, contudo, observa, com base na visão reducionista de explicações neuro-cognitivas proposta por Bickle, que há dificuldades para as visões neomecanicistas. Segundo ele, ao menos em alguns casos, as partes de um mecanismo e seu modo de organização podem servir para explicar o comportamento do mecanismo como um todo. Assim, explicações neomecanicistas se aproximam do reducionismo que seus proponentes tentam evitar, e por isso não conseguem preservar a autonomia das explicações em ciências cognitivas. Leite oferece assim uma elucidação do debate entre a concepção mecanicista e a concepção reducionista de explicação nas ciências cognitivas.

Diana Couto, no texto "(Re-)Interpretando "Thought & Talk: Donald Davidson acerca das mentes animais", propõe uma crítica à interpretação tradicional de Davidson, segundo a qual, para ele, o pensamento dependeria da linguagem. Segundo essa interpretação, ele negaria que criaturas que não tenham linguagem, como animais e bebês, possam ter pensamentos e crenças. Para muitos, essa é a visão que Davidson propõe em textos como o artigo "Thought and talk" (1975). Várias críticas foram levantadas contra Davidson, muitas com base em investigações empíricas sobre seres não-linguísticos. Muitos aceitam que a atribuição de crenças a criaturas não linguísticas explica e prevê seus comportamentos, o que torna útil e justifica pragmaticamente essa atribuição. Davidson também aceita essa ideia mas, segundo Couto, para ele isso não demonstra que seres não-linguísticos de fato possuem crenças. Essa linha de raciocínio leva a autora a propor uma interpretação diferente da posição de Davidson em relação às mentes animais. Segundo ela, Davidson nem afirma nem nega que seres não-linguísticos possuem pensamentos ou crenças, adotando assim uma postura cética. Isso se dá devido à drástica indeterminação explicativa de seus comportamentos. Para Davidson, não se pode caracterizar de forma confiável o que criaturas não linguísticas pensam, dada a sua incapacidade de manifestar respostas verbais. A crítica de Couto ao que intitula de 'Leitura Forte' de Davidson se apoia principalmente em certas concepções de Davidson acerca da intensionalidade e do holismo das crenças, bem como em sua concepção do conceito de crença. Segundo ela, essas teses não implicam que Davidson aceite a tese de que criaturas não-linguísticas não pensam.

No artigo "Reflexive rules as content: the case of deictic demonstratives", Eduarda Calado Barbosa discute um caso de comunicação linguística envolvendo o demonstrativo dêitico "esta". Eis o caso: Jane, que é assistente de dentista, está em seu trabalho e ouve uma conversa entre duas pessoas que aguardam na sala ao lado. Em certo momento, uma dessas pessoas diz "Pai, esta é Julie!". Cabe

aqui salientar alguns aspectos que conferem um caráter não-paradigmático a essa situação: primeiro, Jane não está na sala onde a conversa ocorre, o que significa que ela não tem acesso visual ao ambiente no qual a sentença "Pai, esta é Julie!" foi enunciada. Em segundo lugar, Jane não participa da conversa - ela encontra-se na posição de alguém que ouve o que outros estão conversando sem no entanto ser reconhecida por eles -, o que significa que os participantes da conversa não têm motivos para se preocupar com o que ela compreende ou deixa de compreender. Na prática, eles não vão reiterar ou complementar suas colocações verbais em atenção à Jane. O caso, que não é excepcional na vida cotidiana, configura uma situação instigante: Jane é uma intérprete competente da língua na qual a sentença em comento foi enunciada, mas não participa do cenário específico no qual esse enunciado ocorreu. Segundo Barbosa, mesmo sem conseguir determinar a referência do termo 'esta', Jane apreende algo do enunciado "Pai, esta é Julie!", o que sugere que o papel do demonstrativo dêitico não é meramente indicativo. Qual é o conteúdo expresso por um demonstrativo dêitico quando sua referência não pode ser determinada? Com base na teoria reflexivo-referencial de Perry, Barbosa defende que trata-se de um conteúdo que é sobre o próprio enunciado, e as informações linguísticas que ele carrega. Concretamente, segundo Barbosa, do enunciado "Pai, esta é Julie!", Jane pode apreender que "O indivíduo salientado pelo autor do enunciado 'Pai, esta é Julie!', estando a uma distância *d* do autor do enunciado, chama-se 'Julie'".

No artigo "So language. Very prescribe. Wow.", Shane Nicholas Glackin sugere que um fenômeno recente - os memes "doge", que consistem na foto de um cachorro Shiba Inu com algumas frases de uma ou duas palavras cada - lança luzes sobre o clássico debate entre chomskianos e wittgensteinianos acerca da natureza da linguagem e de como abordá-la cientificamente. Chomsky e seus seguidores consideram que a visão popular segundo a qual a linguagem é algo público e coletivo não tem lugar em uma abordagem científica. Para eles, Glackin explica, "a linguagem é uma característica do cérebro de um indivíduo capaz de falar". Consequentemente, o estudo científico dela "deve procurar descrever a língua-I - isto é, as características relevantes do cérebro e da mente de indivíduos falantes - bem como os comportamentos [linguísticos] resultantes, e se abster de tentar avaliar se tais comportamentos são corretos ou não" (Glackin, neste volume - tradução livre). Ora, se a linguagem é uma propriedade de mentes individuais e não de comunidades, então realmente não faz sentido ajuizar que alguém está falando correta ou incorretamente. Por outro lado, para wittgensteinianos, a atitude normativa que as pessoas tipicamente têm para com a linguagem é um indicativo suficiente de que a linguagem ela mesma é normativa e que as regras são públicas. Isso posto, Glackin volta-se para os usos dos memes "doge" e defende que eles constituem evidência da natureza normativa da linguagem. É importante notar, em primeiro lugar, as características da linguagem encontradas nesses memes. Glackin explica que "um enunciado doge típico combina pelo menos duas ou três frases de duas palavras cada, junto com ao menos uma interjeição, usualmente "wow". Tipos híbridos, como "such wow" e "very excite" também são permitidos" (Glackin, neste volume - tradução livre). À primeira vista, a linguagem doge parece não-gramatical ou primitiva. No

entanto, linguistas interessados no fenômeno têm apontado que ela é "construída a partir de uma gramática muito específica que os usuários não seriam capazes de lançar mão sem um conhecimento relativamente sofisticado da gramática oficial da língua inglesa" (CHIVERS, 2014, citado por Glackin). Para Glackin, isso significa que a linguagem doge segue regras, no sentido wittgensteiniano. Mais: falantes competentes dessa linguagem policiam uns aos outros quanto ao seguimento de tais regras em novos memes. Glackin descreve esse fenômeno para pressionar a tradição chomskiana: se essa normatividade inerente à linguagem doge não é uma característica geral da linguagem, então o que faz com que casos como doge sejam especiais, uma vez que não há como descartá-los como sendo não-linguísticos? Contra a tradição chomskiana, Glackin conclui que "normas prescritivas são uma característica real e ubíqua da linguagem e um objeto de estudo real e legítimo para linguistas" (Glackin, neste volume - tradução livre).

No artigo "Análise da *epíkepsis tôn onomáton* de Antístenes", Joedson Silva Santos reconstrói elementos-chave da filosofia lógico-linguística de Antístenes, pensador antigo que foi seguidor de Sócrates. Santos explica que ao lado de Platão e Demóstenes, Antístenes figura "como um dos melhores expoentes do simples e puro estilo ático; adjacente a Platão e Xenofonte, ele é apontado como escritor de habilidade precisa, possuidor de técnica de expressão e de boa reputação". Santos concentra-se no método de análise de termos de Antístenes, uma vez que a investigação dos nomes é chave para entender outros temas da filosofia desse pensador. Em síntese, o método de análise antistênico considera que "os nomes podem ser agrupados em três processos: epíkepsis tôn onomáton - investigação dos nomes - khrêsis tôn onomáton - uso dos nomes - e dialegein katá géne - distinção em classe" (Santos, neste volume). Antístenes confere uma finalidade educativa ao primeiro desses processos, pois considera que a utilização correta dos nomes é fundamental para a apreensão da realidade e, sob esse prisma, é um "princípio ou fundamento da formação intelectual" (MÁRSICO, 2014, p. 259 - citada por Santos). Quanto ao uso dos nomes, Antístenes comunga da crença em uma linguagem objetiva conectada de modo não polissêmico à realidade. Segue, desse entendimento, que o discurso sábio é unívoco, isto é, designa cada coisa pelo nome que lhe é próprio. Assim, em grandes linhas, a filosofia lógico-linguística desdobra-se em questões éticas e pedagógicas.

Em seu texto, “A discussão em torno da ‘parte II’ (TS 234) das *Investigações Filosóficas* de Wittgenstein”, Filicio Mulinari investiga as razões históricas que levaram os primeiros editores das *Investigações Filosóficas* a considerar escritos da fase tardia da filosofia de Wittgenstein como continuação da obra. Em edição mais recente, o texto presente na tradicional ‘parte II’ é considerado como a *Filosofia da Psicologia* de Wittgenstein. Segundo Mulinari, apesar das evidências históricas não serem suficientes para definir a questão, alguns comentadores de Wittgenstein defendem que o conteúdo do texto revela uma terceira fase em seu pensamento, o que enfatiza a ideia de que a chamada ‘parte II’ não seria uma continuação do texto principal das *Investigações*

Filosóficas. Por outro lado, afirma o autor, outros comentadores defendem e apresentam evidências textuais de que a parte II é uma continuação da primeira parte das *Investigações Filosóficas.* Mas essas evidências abrem margem para interpretarmos outros textos da obra de Wittgenstein como parte das *Investigações Filosóficas*. Apesar disso, devido a outros fatores, o autor afirma que essas evidências não são suficientes para tirarmos conclusões a respeito do pertencimento ou não da parte II à obra principal. Mulinari finaliza afirmando que o processo de escrita das *Investigações Filosóficas* como um todo se deu como que em camadas, pois inclusões e alterações eram feitas a cada revisão. Devido a essa e demais características do texto, ambas as perspectivas, a saber, tanto que a filosofia da psicologia é continuação da obra principal quanto que ela é um trabalho distinto podem ser sustentadas.

Embora a abordagem de Mulinari seja majoritariamente histórica, a discussão do texto nos leva a uma questão mais geral: como a filosofia da mente se relaciona com a filosofia da linguagem? Como mente e linguagem interagem? Certamente, a linguagem importa para a filosofia da mente e a mente importa para a filosofia da linguagem, mas ainda há muitas questões em aberto. Este dossiê oferece alguns recortes e desdobramentos de questões centrais acerca da mente e da linguagem, considerando não apenas a filosofia contemporânea, mas também oferecendo espaço para a discussão de obras históricas. Todas elas prometem iluminar e incitar o debate filosófico sobre as questões centrais que o volume se propõe a abordar.

Agradecimentos

Nós, os editores do Dossiê Filosofia da Mente e da Linguagem, gostaríamos de agradecer primeiramente ao Editor-Chefe e à sua equipe pelo apoio e disponibilidade no decorrer do processo editorial.

Agradecemos e reconhecemos o valor do trabalho dos pareceristas, sem o qual não seria possível a seleção e o aprimoramento dos textos que compõem este volume.

Ficamos felizes com o acolhimento da proposta do dossiê tanto pela comunidade acadêmica nacional quanto pela internacional, por meio de submissões de qualidade e pareceres cuidadosos.

Cumprimentamos e agradecemos a todas as pessoas que submeteram seus trabalhos para avaliação. Recebemos um grande número de textos e, infelizmente, muitos não puderam ser incluídos. Esperamos que esses textos possam encontrar em breve seus lugares aos olhos do público.

Nossos cumprimentos especiais e agradecimentos aos autores dos textos selecionados.

E, por fim, mas principalmente, desejamos que você, leitor, possa usufruir deste volume e fomentar o debate filosófico a partir das contribuições que os autores aqui apresentam.

THE EXTENDED MIND HYPOTHESIS: AN ANTI-METAPHYSICAL VACCINE

A HIPÓTESE DA MENTE ESTENDIDA: UMA VACINA ANTI-METAFÍSICA

GIORGIO AIROLDI[1]
Universidad Nacional de Educación a Distancia (UNED) – Spain
gairoldi1@alumno.uned.es

ABSTRACT: Discussions about the extended mind have 'extended' in various directions in the last decades. While applied to other aspects of human cognition and even consciousness, the extended-mind hypothesis has also been criticized, as it questions fundamental ideas such as the image of a dual world, divided between an external and an internal domain by the border of 'skin and skull', the idea of a localized and constant decision center, and the role of internal representations. We suggest that the main virtue of the hypothesis is not as a theory per se, but as a vaccine against persistent metaphysical prejudices about the mind's structure, functions and borders. Being an hypothesis about the most efficient ways to combine resources and problems, and not a theory about the mind's a-priori constitution, the extended mind view moves the focus from ontology to pragmatics and helps purify philosophy of mind from metaphysical remainders.

KEYWORDS: Extended mind. Externalism. Internalism. Causal-constitutive fallacy.

RESUMO: *As discussões sobre a mente estendida "se estenderam" em várias direções nas últimas décadas. Embora aplicada a outros aspectos da cognição humana e até à consciência, a hipótese da mente estendida também tem sido criticada, pois questiona ideias fundamentais como a imagem de um mundo duplo, dividido entre um domínio externo e um interno pela fronteira da "pele e do crâneo", a ideia de um centro de decisão localizado e constante, e o papel das representações internas. Sugerimos que a principal virtude da hipótese não é uma teoria per se, mas uma vacina contra preconceitos metafísicos persistentes sobre a estrutura, funções e fronteiras da mente. Sendo uma hipótese sobre os modos mais eficientes de combinar recursos e problemas, e não uma teoria sobre a constituição a priori da mente, a visão da mente estendida move o foco da ontologia para a pragmática e ajuda a purificar a filosofia da mente de resquícios metafísicos.*

PALAVRAS-CHAVE: *Mente estendida. Externalismo. Internalismo. Falácia causal-constitutiva.*

INTRODUCTION

Since the publication of "The Extended Mind" (CLARK & CHALMERS, 1998), discussions around the extended mind have flourished and 'extended' in various directions. On the one hand, the extended cognition hypothesis, in what is known as its 'the second wave', has been developed and fine-tuned, and applications of its principle have been proposed for other aspects of human cognition like

[1] Department of Logic, History and Philosophy of Science at Universidad Nacional de Educación a Distancia (UNED).

perceptions, emotions, and even consciousness. On the another hand, radical criticisms of the concepts upon which the hypothesis is based (e.g. the definition of cognitive, its explanatory power, etc.) have led to qualifying and in some cases to reformulating some of its facets. Nevertheless, almost two decades after it was first proposed, the hypothesis seems to be in wonderful shape. This is to be ascribed to the fact that it addresses some of the deepest issues in philosophy of mind, such as the differences between internal and external, the function of representation, and the ontology of mind; and to the fact that, thanks to the diatribes that it fosters, it represents a powerful vaccine against metaphysical temptations and a constant incentive to the progress of research.

In this paper, we present first of all Clark and Chalmers' original idea, in order to illustrate the main hypothesis upon which it rests. In the second section, we briefly show the radical differences between the focus of the first and the second wave of arguments. The third section presents some of the proposals of extension of the hypothesis to other areas of cognition. In the fourth section, we propose a classification of the different stances about the mind that oppose each other in various degrees (internalism, embodied and embedded perspectives, etc.). Finally, we list the main discussions that arise with respect to the causal-constitutive fallacy, the mark of the cognitive, the difference between internal and external processes, and the explanatory power of the hypothesis of extension. The paper finishes with some reflections and conclusions around what we consider to be the virtues of the hypothesis of the extended mind and the reasons for its success.

1 Internalism, externalism and the extended mind

The extended mind hypothesis is a development of externalism. According to externalism, the contents of intentional states are relational properties identifiable only with reference to some reality (physical, linguistic, social, or metaphysical) external to the individual. On the opposite side, internalism considers intentional states as intrinsic properties of the individual, independent from the external reality (BUZZONI, 2006).[2] The 'classical' externalism initially proposed by Putnam (PUTNAM, 1976) and Burge (BURGE, 1979) does not, nevertheless, grant any active role to external elements in the determination of the intentional content: these play a purely passive role in cognitive processes and in the definition of the mental states of the individual.

Clark and Chalmers' extended mind hypothesis introduces a new type of externalism, in which external elements do have an 'active' causal role (CLARK & CHALMERS, 1998).[3] The central issue in the Otto case (CLARK & CHALMERS, 1998)[4] is not to counter a 'broad' versus a 'narrow' content, as it would be the case with

[2] The debate between internalism and externalism was initially fostered by the thought experiment of Twin Earth, proposed by Putnam (1975). Burge (1979) extends the externalist hypothesis and claims that mental content depends upon the social environment. For an overview of linguistic externalism see Bezuidenhout (2008).
[3] "[...] the relevant external features are *active*" (CLARK & CHALMERS, 1998, p. 09).
[4] In the article, the case of Otto is described, a patient with Alzheimers who stores his beliefs in a portable notebook that he always keeps with himself.

classical externalism, but to defend a narrow content extended to the environment (CLARK, 2010a).[5] An example of cognitive phenomenon of this kind is the act of gesturing while talking, an activity which is corporal and neuronal at the same time: verbal thoughts and physical gestures influence each other and constitute a coupled system. Goldin-Meadow (2003) suggests that gesture can provide an alternative representational format that adds information in either an analog, motoric, or visuospatial way. Additionally, gesture can reduce the overall neural cognitive load and free resources for other tasks.[6] Gestures, thus, do not just express internal thoughts complete by themselves, but are part of these thoughts and causally interact with the cognitive system (CLARK, 2011, p. 123-126).

The heart of the issue here is not the trivial observation that external elements causally influence the process (for example, by increasing our memory or by helping us when making calculations). Clark and Chalmers affirm that the causal relationship between external and internal elements, far from being distant and historical, is mutual and diachronic. Clark applies the notion of 'continuous reciprocal causation' to cognitive activities that involve "continuous, mutually modulatory influences linking brain, body, and world" (CLARK, 1997, p. 163): playing in an orchestra or having a group conversation are examples of such activities whose explanation cannot be given in terms of input to and output from a closed cognitive system (CLARK, 1997, p. 165). The cognitive system consists of the sum of both kinds of elements (external and internal), and constitutes a real coupled dynamical system (MENARY, 2010a, p. 03-04).

Not all coupled dynamical systems are cognitive, though. Clark and Chalmers propose a fundamental criterion to limit the systems that can be really considered 'extended minds': the Parity Principle. The principle affirms that an element[7] belongs to a cognitive system not because of its localization (for example, within or outside of the brain), but because of its function. The authors list portability, availability, typical and uncritical use, and easiness of access as necessary and sufficient characteristics for an element to belong to the cognitive system: according to these criteria, a notebook would not count as such, while Otto's portable notebook would (CLARK, 2010a, p. 44-47). The Parity Principle clearly shows the commitment of the original extended mind hypothesis with functionalism[8], as it identifies and classifies mental states in terms of their causal roles (BECHTEL, 1988).[9] Together with functionalism's flaws,[10] the extended mind hypothesis shares its main virtue: it dissociates the cognitive from the physical; a silicon circuit or a Martian organ can be as cognitive as a human neuronal circuit.

[5] "[...] what was at issue was more like an environmentally extended case of narrow content than a case of broad content" (CLARK, 2010a, p. 45).
[6] Goldin-Meadow (2003) presents the case of reduced capacity of memorizing a list of words by a group of children not allowed to gesture with respect to another group which could freely gesture.
[7] The reference to 'an element' is important, given that the principle is a criterion of 'belonging to' and not of constitution: distinction that, as it will be shown, lies at the basis of the discussion with Adams and Aizawa.
[8] The 'second wave' of the extended mind hypothesis overcomes this dependency.
[9] "Functionalism maintains that mental events are classified in terms of their causal role" (BECHTEL, 1988, p. 112).
[10] Bechtel (1988, p. 123-136) lists a summary of the main criticisms to functionalism.

There are different views with regards to the reach of this functionalism: Clark considers that the Parity Principle involves a 'very weak' functionalism that, according to Chalmers (CLARK, 2011),[11] does not extend to consciousness, and that just denies any relevance of the internal/external difference for the cognitive processes; Wheeler defends, on the contrary, an extended functionalism (WHEELER, 2010).[12]

The classical vision considers the mind as a Turing machine with a certain ability for computation and manipulation of symbols, result of the accumulation of adaptive responses. The extended mind hypothesis suggests an idea of human cognition radically different, as a product of the hybridization between the brain and technological artifacts, implying that human beings are 'natural cyborgs'. There is a continuous range of intermediate cases between behaviors and decisions based on discursive rationality and 'quasi-automatic' ones, in which the active contribution of the body, of the social environment, and of cultural artifacts can sometimes prevail over the role of the brain in solving problems, freeing the latter from a relevant working load. The extended mind hypothesis is about reinventing cognition as a distributed capacity (CLARK, 2001, p. 121-129).

Under this view, human cognition springs from the collaboration among body, brain, and the active contribution of the technological environment: the artist draws a sketch before painting the landscape because the sketch is part of the creative process as much as her hand or brain, and not just a simple temporary information store (CLARK, 2001, p. 133).

2 The two 'waves' of the hypothesis: parity and complementarity

The internalist view considers that the brain is the only place where cognitive activities take place, and attributes to this organ the capacity of unplugging from the environment and of managing representations. The extended mind hypothesis rejects this possibility by highlighting two mechanisms at the basis of the extension: functionality and complementarity.

The Parity Principle suggests that it is the function, and not the location, what makes something cognitive: an external element belongs to the cognitive system because it plays a functional role identical to the one that an internal element would play. The principle imposes a 'veil of ignorance' that makes irrelevant the border of 'skull and skin' and the difference between perception and introspection, and thus avoids the 'bio-chauvinist' prejudice: portability and availability of a resource are the only fundamental virtues in cognition (CLARK, 2011, p. 78). By assuming an isomorphism between internal and external processes, however, the Parity Principle is unable to capture the differences between *exograms* (external symbols) and *engrams* (cerebral memory):[13] a

[11] "I think that functionalism about consciousness is implausible" (CHALMERS, 2011, p. xv).
[12] "[…] *the parity principle* forges a strong connection between functionalism and ExC [extended cognition hypothesis]" (WHEELER, 2010, p. 248).
[13] Terminology suggested by Donald (1991, p. 314), who names *exograms* after Lashley's (1950) *engrams*, single entries in the biological memory system.

checklist and an engram are not equivalent in many ways; moreover, it does not distinguish individual differences with regards to the use of available resources (for example, some people might prefer memorizing information, while others checking it on an agenda).

The Complementary Principle (SUTTON, 2010, p. 194), that opens up the so-called 'second wave' of the extended mind hypothesis, overcomes these problems by including an external element in the cognitive system precisely because it plays a different role than an internal one, and that the latter could not perform. By focusing on the cognitive contribution of the specific features of non-neuronal elements, the principle acknowledges that external traits are not isomorphic with internal ones, but complementary: their substantial difference is precisely what gives value to their contribution (SUTTON, 2010). Heersmink (2015) suggests to identify the kind and level of integration between agents and artifacts on the basis of several dimensions, among which: the direction of the flow of information between the elements of the cognitive system (e.g. road signs are one-way from artifact to agents, post-it notes are two-way from agent to artifact to agent); the reliability of the access to external information (e.g. analogue notebooks are more reliable than electronic ones because they do not require electricity); the durability of the relationship with the artifact (e.g. shopping list are on-offs, abacuses are re-used); the trust granted to the information (e.g. to a Encyclopedia Britanica article vs. a Wikipedia entry); the procedural and informational transparency of the cognitive artifact (i.e. the effortlessness in using and interpreting the artifact); the individualization or interchangeability of the artifact (e.g. a underlined book are not interchangeable with other copies); the transformation of the representational and cognitive capacities of the agent by the use of the artifact (as in enculturation, e.g. the brain of a baby modified when learning to speak).[14] Heersmink do not consider that these dimensions constitute necessary and sufficient conditions: they rather "provide a toolbox for investigating the degree and nature of the integration of agent and artifact into 'new systemic wholes'" (HEERSMINK, 2015, p. 596).

The two approaches (functionalism and complementarism) are not necessarily opposed in all their aspects (KIVERSTEIN & FARINA, 2011), as, some might argue, functional isomorphism is not explicitly required by the principle, as Clark and Chalmers only claim that "[i]f, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain't (all) in the head!" (CLARK & CHALMERS, 1998, p. 08).

[14] Heersmink (2015, p. 591) makes the interesting comment that "this is the reversed version of the parity principle", as "external states and processes, rather than being functionally isomorphic to internal ones, are soaked up by the brain which then takes on the properties of those external representational systems".

3 *EXTENDING* THE EXTENDED MIND: FROM BELIEFS TO CONSCIOUSNESS AND TO NICHE CONSTRUCTION

The extended mind hypothesis considers that external elements not only play a causal role in intelligent behavior, but also show a constitutive inter-dependence with it and, in some cases, allow behavior that would not be possible if the cognitive system lacked them (WHEELER, 2010, p. 246-247): mental processes and beliefs can constitutively include environmental resources. The natural extension of the hypothesis points to the possibility that also other intentional mental states, reasoning, perceptions and emotions could extend to the environment.

Clark claims that such an extension is real: for example, when one uses the camera of a mobile phone as extension of her visual capacity, provided the camera fulfills the criteria of the Parity Principle. Chalmers, consistent with his rejection of functionalism in the realm of consciousness, excludes from it also the extended mind hypothesis, given that he does not conceive a thought experiment in which two identical twins can have different self-experiences: he imputes such impossibility to the band-width required by consciousness to access information, band-width not supported by our current perceptual system (CLARK, 2011, p. xiv-xv). Rowlands, on the contrary, claims that consciousness too can be extended to the environment, moving the focus from the definition of consciousness as intentional object of an apprehension act to consciousness as determinant of a reference (ROWLANDS, 2010).

By denying the possibility of an extended consciousness, and at the same time admitting the possibility of extended states of mind, one grants to the brain a special role, and at the same time rejects the identification of the mental with the conscious.[15] Clark solves this apparent dilemma by diluting his externalism with some dose of internalism, given that he keeps within the brain, besides consciousness, many of the functions that internalism traditionally grants it, such as the ability to create representations and the possibility to develop cognitive processes isolated from the environment (that is, off-line, or non-extended): in short, he accepts the internalist view of the brain as the center of the cognitive. However, conversely to classical internalism, Clark claims that the brain, even if it is the center of the cognitive system, does not bother for the localization (either internal or external) of its resources and processes, as long as they are useful to solve problems. The human cognitive system is 'organism centered', but not 'organism bound'. Clark also clearly rejects the internalist view of the brain as 'final decider': internalism does not specify which part of the brain, and based on which functions, makes decisions, and it is even doubtful that decisions as such appear in a single point of the cognitive system (CLARK, 2011, p. 159-160).

[15] Clark underscores this repeatedly: "Finally, we allowed that conscious mental states [...] supervene only on local processes inside the head. But insofar as the scope of the mental is held to outrun that of conscious..." (2010a, p. 45; 2011, p. 79); "Shrinking the mind to the conscious is certainly one way to avoid the conclusions of the original paper. But do we really want to shrink the mind so far?" (2011, p. 161).

A more radical interpretation considers the extended mind hypothesis as a particular and extreme case of a general and complex hybridization between living beings and environment (physical, social, technological and cultural). The human being is *wholly extended* into the environment, as it is the case with digestion, linked to artifacts and techniques for the preparation and cooking of food. In the *niche construction model* (STERENLY, 2010), agents modify their environment and are modified by it: many animals build dens, and some trees modify the land where they grow (for example, eucalypts increase the acidity of the ground close to their trunk in order to avoid other plants to take root close to them). Niches are built around certain resources and identified with regards to three axes: the availability and reliability of the resource (external resources, for example, being shared, could be more reliable than internal ones); its standardization (external personalized resources, such as Otto's portable notebook, or a blind person's stick, are placed in one end of this axis, given that they are perceived as part of the body; a book full of hand-notes is placed in an intermediate position, as it is personalized but not incorporated); its shared and collective nature (the mnemonic resources of Elizabethan actors are an example of external shared memory (SUTTON, 2010, p. 208-209)). The extended mind hypothesis is, under this view, just a particular and extreme case of niche construction: the case of highly reliable, incorporated and personalized external resources.

4 Degrees of extension

Between pure internalism and active externalism, there are several postures that differ with regards to the extension of the mental and its relationship with the environment: we briefly list the most interesting features of the main ones.

Pure internalism exists in two fundamental versions: one, defended for example by Adams and Aizawa (2001)[16] and defined as *brainbound* by Clark (2011, p. xxvii), limits cognition to the neuronal; the other, known in the literature as *embodied*, recognizes in its radical version a cognitive and explanatory role also for the body (thus approaching externalism) and, in more moderate versions, to representations, especially to mental representations of the body (ADAMS & AIZAWA, 2001).

A 'weak' externalist position, known as the *embedded hypothesis* (RUPERT, 2004) admits that part of the environment has a causal role in some cognitive processes, although it is not a constitutive part of them (Wheeler 2010: 246). The embedded hypothesis highlights the difference between external and internal states (for example, between biological memory and external memory, CLARK, 2011, p. 112-113) to escape the causal-constitutive fallacy[17] (ADAMS & AIZAWA, 2010, p. 67).

[16] "We maintain that […] processes that are plausibly construed to be cognitive occur within the brain, do not occur outside of the brain, and do not cross the bonds of the brain" (ADAMS & AIZAWA, 2010, p. 74).
[17] See section 5 for an analysis of the causal-constitutive fallacy.

Finally, the 'strong' externalist position, linked to the extended mind hypothesis, considers that environment to be a constitutive part of the cognitive system, that in turn includes the brain, the body and the physical, cultural and social environment, all interconnected through systemic relationships of feed-back, feed-forward, etc. (CLARK, 2011, p. xxviii).

The extended mind hypothesis can be taken to the extreme of denying that the brain generates representations. According to this view, the cognitive system resorts to the environment as a model of itself, without any need of creating internal images. This view derives from Brooks' suggestion to move from approaches in Artificial Intelligence highlighting abstract manipulation of symbols to a methodology emphasizing ongoing physical interaction with the environment, provided that "the world is its own best model" (BROOKS, 1999, p. 115): if representations are grounded on the physical world, the need for traditional symbolic representations fades entirely. The known incapacity of the subject, if focused on some detail of a scene, to detect even very relevant changes in other parts of it, is considered as an evidence of this claim: if an internal image existed, any change in the world would be detected, once reality were matched with the mental model. As Merleau-Ponty (2012) points out, the body is always present and is thus the best model of itself. The reason for this rejection of representations is that they are seen as a return to mentalism, as they weaken the causal role and explanatory force of the external element (ALSMITH & DE VIGNEMONT, 2012).[18] There are hypotheses, though, that avoid this risk without coming to such an extreme conclusion (SIMONS & RENSINK, 2005): the representation could, for example, arise and decay quickly, or persist without being taken into account. As a result, the phenomenon of 'blindness to details' might imply not that there are no representations of the environment, but rather that there are several, partial ones, not so easily available to conscious access. Moreover, for the principle of minimum effort, the subject would rather access directly the environment, when available, to gain information, even if an internal representation of it were available (CLARK, 2011, p. 141-146).

The following table summarizes the degree of 'external extension' of the cognitive system defended by the discussed postures: internalism and the extended mind hypothesis just represent the two extreme positions of a rich range of options.

[18] "[...] whether positing body representations actually undermines the explanatory role of the body" (ALSMITH & DE VIGNEMONT, 2012, p. 02).

Brain	Body		Environment	
	Representation	Real	Causal role	Inter-connected role
Internalism/ Brainbound				
Moderate Embodied				
Radical Embodied				
Embedded or Scaffolded				
Extended Mind				

Table 1 – Degrees of extension of the cognitive system

5 Fruitful controversies

The great amount of discussions and fruitful controversies that the extended mind hypothesis has not ceased to foster since its appearance in 1998 is perhaps the best evidence of its importance for philosophy of mind and of the topicality of the issues that it addresses. The main ones are presented in this section.

The causal-constitutive fallacy

Adams and Aizawa defend an internalist position that locates all cognitive processes within the brain, and reject the extended mind hypothesis by denying that it is sufficient for an external element to play a causal role in such processes to be part of them: believing it means falling in the causal-constitutive fallacy (ADAMS & AIZAWA, 2010, p. 73-76).

Answers to this criticism are numerous and assorted: Clark and Menary discard it; Palermos recognizes that it does have some ground and suggests, to escape it, a modification of the extended mind hypothesis; Ross and Ladyman claim that it makes no sense at all. Let's briefly go through each of these positions.

Clark claims that the issue is not to know whether an element is *in itself* cognitive when used by a cognitive system (or one would fall in the 'compositional fallacy', the assumption that the parts of a system must have the same characteristics than the system itself). The question is rather whether an element is part of a cognitive system (CLARK, 2010ab, p. 82-85). Adams and Aizawa's position presupposes that there is a well-defined cognitive agent to which an external

element couples just because it has a casual impact on it: but such an agent does not exist, and the cognitive system is constituted by the whole set of all elements intervening in it. The definition of 'extended' itself might perhaps suggest the idea of a cognitive center: 'distributed' mind would be a more appropriate definition (MENARY, 2010b, p. 606-610).

Palermos accepts that the original version of the extended mind hypothesis exhibits the causal-constitutive fallacy (PALERMOS, 2014).[19] He proposes to apply the theory of dynamical systems[20] to the modelling of cognitive processes in order to avoid the identification of causal and constitutive, and to supplant it with the stronger and more defensible identification of dynamic and non-linear inter-causality. In this way, he claims that the extension hypothesis is saved and the fallacy avoided.

Ross and Ladyman reject discussion about the fallacy because they consider it to be based on metaphysical principles. Questions about the inclusion of an element in a system because of its causal role are senseless, given that neither the concept of system, nor the concepts of causality or constitution have a place in a mature science (ROSS & LADYMAN, 2010).

The mark of the cognitive

The abundance of intermediate postures between internalism and the extended mind hypothesis (Table 1) is in part due to a lack of a universal definition of cognition. Each view develops around a more or less implicit interpretation of what mind is: it is therefore necessary to make it explicit, to avoid using a proposal about the limits of the cognitive system to deduce a definition of cognition that lays at the basis of the initial proposal, plunging into the pitfall of a *petitio principii*. If cognition includes activities such as remembering, perceiving, learning, or reasoning, what do they all share that makes them cognitive? The range of answers is very wide, from the classical computationalism that identifies cognitive vehicles with symbols equipped with syntactic properties; to connectionism that identifies them with algorithms of nodes' activations in a network. In a different line, Wilson (2010, p. 183) moves the focus from representation to the act of representing as a mark of the cognitive, and Rowlands identifies a cognitive process in whatever allows the performing of a cognitive action. Alternatively, the cognitive integration approach acknowledges that our primary cognitive engagements with the world are embodied and primarily sensorimotor ones. It thus integrates the bodily 'internal' and 'external' facets of cognition, and understands this integration in terms of manipulation of environmental vehicles (MENARY, 2010a, p. 268-269). The integrationist view includes the linguistic and representational environment within the cognitive system, given that they allow cognitive actions impossible based on neural activity alone (MENARY, 2010b, p. 611). Sometimes, the distinction between the mental and the cognitive, and how the extended mind hypothesis is reflected

[19] "Such a 'cognitive bloat' would actually be the outcome of repeatedly committing the 'causal-constitution' fallacy that Adams and Aizawa pointed out" (PALERMOS, 2014, p. 07).
[20] See next section for a brief description of the theory.

in each, are also poorly defined. Carter and colleagues (CARTER et al., 2014), for example, do explicitly distinguish among extended *cognition*, or the claim that *cognitive processing* can be suprapersonal, extended *mind*, or the claim that *mental processes* can be suprapersonal, and distributed cognition, or the claim that cognitive processing can be distributed across several agents and artifacts. In view of such variety, let's review the main postures regarding the mark of the cognitive.

Adams and Aizawa reduce the cognitive to the manipulation of representations with 'non-derived' content, located uniquely in the brain. What is external to the brain, processes and representations, has always a derived content (for example, the meaning of alphabetical signs derives from social norms): evidence of this is that we can identify an external representation and socially agree to change its content (for example, that the meaning of green for traffic lights becomes 'go'), but we cannot do the same with internal representations (because we do not know which neuronal networks identify them,[21] and we would anyhow not have the ability to manipulate them)[22]. Aizawa and Adams suggest that "there is nothing in DNA or its causal activity during development that resembles the way that meaning is assigned by a human mind to an artifact or Symbol" and that "the derivation of the human mind from the human genome is unlike the derivation of derived content from prior content" (2005, p. 667): it is, therefore, non-derivative. There are many criticisms of this position. In the first place, the authors neither supply an exact definition of 'representation with non-derived content', nor a definition of what differentiates it from one with intrinsic content (CLARK, 2010a, p. 90). Moreover, the proposal seems to be a definition rather than a fact, given that the authors do not justify the view that the cognitive is characterized by intrinsic content: they even claim that it is not clear to what extent cognitive states could also involve derived contents.[23] Dennett denies the existence of original (or non-derived) intentionality by asking "[w]here, though, do we get our 'original' and underived intentionality? From God, as Michelangelo suggests?" (DENNETT, 1990, p. 54).

According to Clark, a process is cognitive if it supports intelligent behavior. The focus is substantially different than in Adams and Aizawa: the mark of the cognitive does not lay in the nature of an element (derived or non-derived content), but depends on its function within the system. Otto's notebook does not (and can not) have any non-derived content, but it is functionally linked to Otto's cognitive system in a dispositional way, in a coupling defined by the Parity Principle. The cognitive is not an intrinsic feature of any isolated element, it rather supervenes on a system whose elements as a set form a cognitive process.

Hurley claims that the putative 'causal-constitutive fallacy' simply masks the prejudice of identifying cognitive with 'internal', while limiting the external role to a causal contribution (HURLEY, 2010, p. 126). This posture considers the internal

[21] "We don't know what specific syntactic item in the brain bears that content" (ADAMS & AIZAWA, 2010, p. 73).
[22] "[…] we have no way to identify particular tokens of brain states qua syntactic items in order to affix contents to them" (ADAMS & AIZAWA, 2010, p. 72).
[23] "It is unclear to what extent each cognitive state of each cognitive process must involve non-derived content" (ADAMS & AIZAWA, 2001, p. 50).

and the external as separate domains with fixed properties: but technology, for example, is not necessarily something external, given that learning a sentence by heart means creating a 'mental artefact' (SUTTON, 2010, p. 207-208). Moreover, there might exist sets of neurons with just a causal role, even if they constitutively belong to the brain.

Butler (1998, p. 205) identifies the cognitive with the biological brain, because it is where the computational control happens: it rejects the idea that external processes could be cognitive. There are several answers to this view: Clark criticizes it first of all because it does not clarify where exactly decisions would be made; on the other hand, there are zones of the brain that do not take part in the decision process, thus according to this definition they would not be cognitive either; finally, the functional role of the biological memory and of the external memory is identical, so it is unclear why the former would be cognitive and the later would not. Even if a place where the final decision is made existed, it is not clear why it should be the brain (CLARK, 2010a, p. 55-56).

Grush (2003) identifies the mark of the cognitive with the capacity of the brain to control motor activities, and not with its capacity to generate representations, that are just 'working tools'. Control of some motor activities is possible thanks to feed-back from the environment; given that feed-back's perception might experience a delay, the cognitive system produces inner dynamical models of the environment that allow to simulate its answers. Representations are nothing more than these off-line models that are used as surrogates of the real environment. Cognitive systems can thus work at the same time by interacting with the real world and with models of it: mind is dis-engaged from the world, but not dis-embodied from the body.

Palermos (2014) criticizes Clark's Parity Principle as the criterion to define the cognitive. Interpreting it as 'glue and trust' principle,[24] he claims that it allows the inclusion of almost any external element, causing a 'cognitive bloat' and falling in the causal-constitutive fallacy. As an alternative, he suggests an interpretation of the cognitive based on the theory of dynamic systems (DST). Contrary to the sequential computation of the classic theory, in which the temporal factor is irrelevant, DST models the cognitive system as the sum of coupled systems, whose mutual and continuous interactions can be described only through non-linear parameters. The advantage of this proposal is that it entails a concept of the cognitive as supervenient and not linked to any isolated element. Clark seems anyhow not to limit his criterion to the Parity Principle and to suggest, in the same

[24] The 'glue and trust' principle consists in a set of conditions under which cognitive processes should be implemented outside the body in order to represent extended cognition: in particular, that "the resource be reliably available and typically invoked […] [t]hat any information thus retrieved be more-or-less automatically endorsed […] [t]hat the information contained in the resource be easily accessible as and when required" (CLARK, 2010a, p. 46). These are just a set of sufficient but not necessary conditions for extended cognition, as other sets can be proposed (AIZAWA, 2015). The equivalence between the Parity Principle and the 'glue and trust' principle can thus be questioned, but Palermo's criticism is not invalidated by this clarification.

line of Palermo's proposal, the need of a strong interconnection among the elements that constitute the cognitive system (CLARK, 2011).[25]

Finally, Ross and Ladyman do not suggest any criterion of the cognitive, but consider that any aspiration of doing so is metaphysical (ROSS & LADYMAN, 2010). The concept of 'constitutive' belongs to a vision of the world that implies the existence of fundamental elements, out of which all the remaining is composed: it is a vision overcome by mature sciences. The question about the border of the cognitive is to be rejected independently of the answer, because it is an ontological question: even the externalist proposal based on the extended mind hypothesis fosters a vision of the world full of spatially identifiable objects.

In conclusion, the various proposals around the mark of the cognitive as well as the controversies related to the causal-constitutive fallacy, and above all Ross and Ladyman's rejection of any criterion as metaphysical, suggest that the concept of mind itself is superfluous in the cognitive sciences.[26] The predicate 'mental' is applied to such a wide variety of realities (from thermostats[27] to complex systems within human beings), that the suspicion arises we are not able to recognize a mind when we see one (CLARK, 2010a, p. 62-64).

Internal and external processes

Several criticisms to the extended mind hypothesis highlight the differences between internal and external processes.

Dartnall stresses that the internal biological memory, contrary to the external, is not just a simple informational store: remembering is mainly *creating* information. Clark answers that the memory of Otto's notebook is not cognitive in itself, in the same way as a group of neurons is not a group that can likewise store information in a passive way (for example, when we learn by heart a sentence in a language we do not understand) (CLARK, 2010a, p. 52-53).

Butler (1998, p. 211) emphasizes that Otto's access to information requires some perception, while Inga's access to her memory is fully internal.[28] Clark answers that Otto and his notebook constitute a cognitive system *as a set*, therefore the whole process remains within it. Davies (apud CLARK, 2010a, p. 57) adds that Otto could make mistakes while reading, but Clark emphasizes that Inga too could remember wrongly. Finally, the objection that Otto's notebook is public and Inga's memory is private is rejected by Clark by appealing to cases of multiple personalities sharing the same memories (CLARK, 2010a, p. 57-58).

---

[25] "Coupling alone is not enough [...] these are the cases when we confront a recognizably cognitive process, running in some agent, that creates outputs (speech, gesture, expressive movements, written words) that, recycled as inputs, drive the cognitive process along" (CLARK, 2011, p. 131).
[26] "[...] a fully general theory of cognition [...] need incorporate no single overarching account of limits on the boundaries of cognitive systems" (ROSS & LADYMAN, 2010, p. 156).
[27] See e.g. Dennett (1987).
[28] Inga is cited as an example of a person with a memory not affected by Alzheimer (CLARK & CHALMERS, 1998).

Sterelny (2004, p. 246) highlights a substantial difference between perception of the external and introspection: the former can be manipulated. The susceptibility to external attacks differentiates the two kinds of channels. Clark, even if he admits that doubts about the reliability of the information stored in his notebook would cause its decoupling from Otto's cognitive system, claims that the same problem can happen in the biological memory, in which a psychologist could generate false beliefs (CLARK, 2010a, p. 60-61).

Critics of the extended mind hypothesis have also insisted that Otto's access to the information stored in his notebook entails two steps (he believes that his belief is in the notebook and then he retrieves it), while Inga's access to her memory is direct. Moreover, Inga can claim to have a 'first person' authority upon her internal beliefs, while Otto has to find out what it is that he believes, by checking in his notebook (PRESTON, 2010). Clark answers that the use of his notebook has become so automatic for Otto that he does not even notice it anymore, the same as Inga's memory: it's a tool in Heidegger's sense (HEDEGGER, 1996; SCHMITT, 1965).

Finally, Rupert claims that a 'classical' cognitive system is persistent, while an extended one is brief: it only works as long as its external elements are available (RUPERT, 2010). Clark answers that the spider's web is not always available either, but this does not hinder the spider's hunting system to be constituted by both spider and net.

Explanatory power

Another criticism of the extended mind hypothesis points out that it does not provide greater explanatory power than the internalist view; moreover, internalism has allowed much progress in the cognitive sciences that should not be abandoned without a clearly better alternative. Adams and Aizawa advocate for keeping a clear separation between the internal field, where interesting regularities have been identified, and the external one, where the enormous variety of phenomena seems to reduce the possibility of finding general laws (ADAMS & AIZAWA, 2010).

Clark rejects this distinction because he believes that the differences between internal and external processes are not greater than those among internal processes (CLARK, 2010a, p. 51). Moreover, Sutton emphasizes that rejecting the study of external cognitive processes because of their variety is like rejecting the study of the nature of mirrors by focusing on the variety of the images they reflect (SUTTON, 2010, p. 214).

Finally, the request for unification made by Adams and Aizawa is fulfilled, given that the extended mind hypothesis really *unifies* explanations about the working of cognitive systems (CHALMERS, 2011, p. xiv).

Conclusions

The multiplicity of the issues under discussion and of the controversies surrounding the extended mind hypothesis are good witnesses of its fecundity. Rather than in the enunciation of a theoretical hypothesis, however interesting and revolutionary, we believe that the main virtue of the idea that some of the cognitive processes can take place outside of the brain consists in its making explicit and questioning prejudices concerning what the mind is, what functions it performs and how, and where its center and its border are located. Many of these issues revolve around the metaphysical view of mind as a substance. From this perspective, the extended mind hypothesis addresses many of the most fundamental areas of philosophy of mind and acts as a powerful vaccine against the danger of more or less explicit metaphysical infections.

First of all, it shows the persistent difficulty in getting rid of the Cartesian dual world, divided between an external and an internal domain, each characterized by its own and incommensurable laws, where, within the border of 'skin and skull', there resides the 'interior homunculus', who contemplates and manipulates the representations that appear in the 'mirror of nature'[29] that is the mind. The metaphysics of Cartesian dualism, if endorsed, prevents the understanding of the problem of mental cognition and of mental structure (KENNY, 1992). Radical internalist views, such as Adams and Aizawa's, with their doubtful definition of the cognitive as 'representations with non-derived content', still show a close proximity to a quasi-magical idea of the internal: because they expel from the realm of the mental all acts not based on representations and decouple the mind from its environment, that cannot generate them; and because they claim that only internal processes can be classified and formalized in laws, contrary to the chaotic external world, putatively full of dissimilar phenomena. The internalist proposal can be successfully applied in some cases, but it should be considered an hypothesis useful as a theoretical ideal and not as the description of an ontological reality.

The questioning of the centrality of the internal has important consequences in two other aspects of cognition: the existence of a defined and constant center where decisions are made, and the role of representations. Let's see how.

Regarding the belief that there exists a central mind that makes all the decisions (or, at least, the important ones), already questioned by Dennett (1991), the idea of a cognition distributed and extended to the environment compels us to reflect about why the distributed structure should stop at the limits of the brain and be centralized. Even if one claims, as discussed in the previous paragraph, that all the cognitive states and processes reside in the brain, one can still admit that decision-making is distributed. On the other hand, with regards to the putatively fundamental role of representations in cognition, we have seen that the active role of external resources suggests the possibility of automatic rational actions not based on representations, and of cognitive processes that give priority to the direct access to the world, when it is available. The role of internal representations in problem

[29] According to the definition in Rorty (1979), where he defends an anti-representationalist posture.

solving is also doubtful, as long as cognition is active and distributed: representations would play a role similar to the one of external elements (CLARK, 2001, p. 129-131).

Besides, the extended mind hypothesis moves the focus from ontology to pragmatic considerations, as it is an hypothesis about the most efficient ways to combine resources and problems, and not a theory about the *a-priori* constitution of the mind. The Parity Principle is not a definition of the cognitive (as a matter of fact, Clark doesn't engage with any specific definition) and it just identifies a criterion of belonging: instead of establishing first what the cognitive is in a static and ontological way (as if the cognitive were a natural kind)[30], and afterwards deciding whether an element that takes part in a cognitive process fulfills the criteria, Clark defines first of all under which conditions an element can be considered part of a cognitive system, without worrying about its ontological status. This posture seems reasonable, because nothing prevents some neural networks from playing a cognitive role in some cases. If, like in the case of 'blindness to details', the brain, even when an internal representation is available, directly resorts to the environment should this make its task easier, there is no reason why it should worry about the ontological status of a resource, as long as the resource can perform a given function. Thanks to its impartiality, the brain seems to believe less than some authors about the 'exceptionality' of the interior.

Table 2 summarizes and contrasts the central points of the discussions between the internalist views and the extended mind hypothesis.

Internalism	Extended Mind
Internal/external dualism	Continuity between external and internal
Clear border of the cognitive (skull & skin, body…)	Border of the cognitive variable depending on the process
Representationalism	Opportunism: representation or reality depending on convenience
Internal laws can be formalized, external laws cannot	Same complexity/variety of internal and external phenomena
Controlling center localized and stable	There is no controlling center, decisions distributed according to variable criteria
Ontology first	Function first

Table 2- Main features of internalism versus extended mind hypothesis

[30] In the sense of Quine (1969).

Discussions fostered by the extended mind hypothesis thus represent an exceptional vaccine that helps purify philosophy of mind from metaphysical infections. Defining a concept of cognitive or of mind can be helpful, either as an assumption to limit the scope of an investigation, or as a guideline to organizing theories. On the other hand, searching for an absolute definition does not seem to be more justified than the metaphysical love of taxonomies. Under this instrumental view, none of the two analyzed positions is absolutely truer or better than the other: each can represent a useful tool. One should therefore look not for a clash but rather for a fusion between them, an 'ecumenical' attitude that avoids slipping towards suspicious extremes: if, on one hand, the hypothesis of embedded cognition runs the risk of seeing the external as purely instrumental to the internal, on the other hand, the extended mind hypothesis might forget that the center of human cognition is the organism.

The sum of these views helps us to remember that the target of all research is not the creation of theories but rather the understanding of reality, and that reality is always more complex and elusive than any ontological coat or metaphysical hat we may have for it.[31]

References

ADAMS, Fred; AIZAWA, Ken. The bounds of cognition. *Philosophical Psychology*, v. 14, n. 1, pp. 43-64, 2001.

_____. Defending the bounds of cognition. In: MENARY, Richard A. (ed.). *The Extended Mind*. Cambridge: MIT Press, 2010. pp. 67-80.

AIZAWA, Ken. *Extended cognition. Trust and glue, and knowledge*. Text of the talk at the First Annual Extended Knowledge at the University of Edinburgh, April 22-23, 2015. Available at: https://www.academia.edu/12094812/Extended_Cognition_Trust_and_Glue_and_Knowledge. Accessed in: 2019-03-02.

AIZAWA, Ken; ADAMS, Fred. Defending non-derived content. *Philosophical Psychology*, v. 18, n. 6, pp. 661-669, 2005.

ALSMITH; Adrian J. T.; DE VIGNEMONT, Frédérique. Embodying the mind and representing the body. *Review of Philosophy and Psychology*, v. 3, n. 1, pp. 1-13, 2012.

BECHTEL, William. *Philosophy of mind: an overview for cognitive science*. New York, London: Psychology Press, 1988.

BEZUIDENHOUT, Anne L. Language as Internal. In: LEPORE, Ernest; SMITH, Barry C. (eds.). *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press, 2008. pp. 127-139.

[31] I would like to thank the two anonymous referees whose comments and suggestions have contributed to substantially improving upon the earlier version of this paper.

BROOKS, Rodney A. *Cambrian intelligence. The early history of the new AI.* The MIT Press, 1999.

BURGE, Tyler. Individualism and the mental. *Midwest Studies in Philosophy*, v. 4, n.1, pp. 73-122, 1979.

BUTLER, Keith. L. *Internal affairs: a critique of externalism in the Philosophy of Mind*. Dordrecht: Kluwer, 1998.

BUZZONI, Marco. Esternalismo/Internalismo. In: *Enciclopedia Filosofica Bompiani*, Milano: RCS Libri, 2006. p. 3702.

CARTER, J. Adam; KALLESTRUP, Jeseper; PALERMOS, S. Orestis; PRITCHARD, Duncan. Varieties of externalism. *Philosophical Issues*, v. 24, pp. 63–109, 2004.

CHALMERS, David J. Foreword,. In: CLARK, Andy. *Supersizing the mind: embodiment, action, and cognitive extension*, Oxford: Oxford University Press, 2011. pp. ix-xvi.

CLARK, Andy. *Being there. Putting brain, body, and world together again.* The MIT Press, 1997.

_____. Reason, robots and the extended mind. *Mind & Language*, v. 16, n. 2, pp. 121-145, 2001.

_____. Memento's revenge: the extended mind extended. In: MENARY, Richard (ed.). *The extended mind*. Cambridge: MIT Press, 2010a. pp. 43-66.

_____. Coupling, constitution and the cognitive kind. In: MENARY, Richard (ed.). *The extended mind*. Cambridge: MIT Press, 2010b. pp. 81-99.

_____. *Supersizing the mind: embodiment, action, and cognitive extension*, Oxford: Oxford University Press, 2011.

_____; CHALMERS, David. J. The extended mind. *Analysis*, v. 58, n. 1, pp. 7-19, 1998. Reprinted in MENARY, Richard (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 27-42.

DENNETT, Daniel C. *The intentional stance*. Cambridge, MA: MIT Press, 1987.

_____. The myth of original intentionality. In: MOHYELDIN SAID, K. A.; NEWTON-SMITH, W. H.; VIALE, R.; WILKES, K. V. (Eds.). *Modelling the mind*. Oxford, England: Oxford University Press, 1990. pp. 43-62.

_____. *Consciousness explained*. Boston. New York, London: Little, Brown and Co., Back Bay Books, 1991.

DONALD, Merlin. *Origins of the modern mind. Three stages in the evolution of culture and cognition.* Cambridge University Press, 1991.

GOLDIN-MEADOW, Susan. *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press, 2003.

GRUSH, Rick. In defense of some cartesian assumptions concerning the brain and its operations. *Biology and Philosophy*, v. 18, n. 1, pp. 53-92, 2003.

HEERSMINK, Richard. Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, v. 14, n. 3, pp. 577-598, 2015.

HEIDEGGER, Marin. Being and time. Translated by Joan Stambaugh. SUNY Press, 1996.

HURLEY, Susan. Varieties of externalism. In: MENARY, Richard (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 101-153.

KENNY, Anthony. *The metaphysics of mind*. New York/Oxford: Oxford University Press, 1992.

KIVERSTEIN, Julian; FARINA, Mirko. Embraining culture: leaky minds and spongy brains. *Teorema: Revista Internacional De Filosofía*, v. 30, n. 2, pp. 35-53, 2011.

LASHLEY, Karl S. In search of the engram. In: F. BEACH; D. O. HEBB; C. MORGAN; H. NISSEN (eds.). *The neuropsychology of Lashley*. New York: McGraw-Hill, 1950.

MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010a.

_____. The holy grail of cognitivism: a response to Adams and Aizawa. *Phenomenology and Cognitive Sciences*, v. 9, n. 4, pp. 605-618, 2010b.

MERLEAU-PONTY, Maurice. *Phenomenology of perception*. Oxford: Routledge, 2012.

PALERMOS, S. Orestis. Loops, constitution, and cognitive extension. *Cognitive Systems Research*, v. 27, pp. 25-41, 2014.

PRESTON, John. The extended mind, the concept of belief, and epistemic credit. In: MENARY, Richard (ed.). *The Extended mind*. Cambridge: MIT Press, 2010. pp. 359-362.

PUTNAM, Hilary. The meaning of meaning. *Minnesota Studies in the Philosophy of Science*, v. 7, pp. 131-193, 1976.

QUINE, Willard van Orman. Natural kinds. In: *Ontological Reality & Other Essays*. New York: Columbia University Press, 1969.

RORTY, Richard. *Philosophy and the mirror of nature*. Oxford: Basil Blackwell Ltd, 1980.

ROSS, Don; LADYMAN, James. The alleged coupling-constitution fallacy and the mature science. In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 155-166.

ROWLANDS, Mark. Consciousness, broadly construed. In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 271-294.

RUPERT, Robert. D. Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, v. 101, n. 8, pp. 389-428, 2004.

_____. Representation in extended cognitive systems: does the scaffolding of language extend the mind? In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 326-328.

SCHMITT, Richard. Heidegger's analysis of 'tool'. *The Monist*, v. 49, n. 1, pp. 70–86, 1965.

SIMONS, Daniel J.; RENSINK, Ronald A. Change blindness: past, present and future. *Trends in Cognitive Sciences*, v. 1, n. 7, pp. 261-267, 2005.

STERELNY, Kim. Externalism, epistemic artifacts and the extended mind. In: SCHANTZ, Richard (ed.). *The externalist challenge. New studies on Cognition and Intentionality*. Berlin & New York: de Gruyter, 2004. pp. 239-254.

_____. Minds: extended or scaffolded? *Phenomenology and Cognitive Sciences*, v. 9, pp. 465-482, 2010.

SUTTON, John. Exograms and interdisciplinarity: history, the extended mind, and the civilizing. In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 189-225.

WHEELER, Michael. In defense of extended functionalism. In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 245-270.

WILSON, Robert, A. Meaning making and the mind of the externalist. In: MENARY, Richard A. (ed.). *The extended mind*. Cambridge: MIT Press, 2010. pp. 167-188.

Recebido em: 03-03-2019

Aceito para publicação em: 25-06-19

THE OBSCURE CONTENT OF HALLUCINATION

O OBSCURO CONTEÚDO DA ALUCINAÇÃO

MARCO AURÉLIO SOUSA ALVES[1]
Universidade Federal de São João Del-Rey (UFSJ) – Brasil
marcoaurelioalves@ufsj.edu.br

ABSTRACT: Michael Tye (2009) proposed a way of understanding the content of hallucinatory experiences. Somewhat independently, Mark Johnston (2004) provided us with elements to think about the content of hallucination. In this paper, their views are compared and evaluated. Both their theories present intricate combinations of conjunctivist and disjunctivist strategies to account for perceptual content. An alternative view (called "the epistemic conception of hallucination"), which develops a radically disjunctivist account, is considered and rejected. Finally, the paper raises some metaphysical difficulties that seem to threaten any conjunctivist theory and to lead the debate to a dilemma: strong disjunctivists cannot explain the subjective indistinguishability between veridical and hallucinatory experiences, whereas conjunctivists cannot explain what veridical and hallucinatory experiences have in common. This dilemma is left here as an open challenge.

KEYWORDS: Hallucination. Perceptual content. Disjunctivism. Michael Tye. Mark Johnston.

RESUMO: *Michael Tye (2009) propôs uma forma de compreender o conteúdo das experiências alucinatórias. Mark Johnston (2004), por vias um tanto independentes, ofereceu também elementos que nos ajudam a compreender o conteúdo das alucinações. As teorias desses dois filósofos são comparadas e avaliadas nesse artigo. Ambas combinam abordagens conjuntivistas e disjuntivistas do conteúdo perceptivo. Uma teoria alternativa, chamada de "concepção epistêmica da alucinação", que defende uma abordagem radicalmente disjuntivista, é considerada e rejeitada. Por fim, o artigo levanta algumas dificuldades metafísicas que parecem ameaçar qualquer teoria conjuntivista e levar o debate para um dilema: os disjuntivistas radicais não conseguem explicar a indistinção subjetiva entre a percepção verídica e a experiência alucinatória, e os conjuntivistas não conseguem explicar o que a percepção verídica e a alucinação têm em comum. Esse dilema é deixado aqui como uma questão aberta.*

PALAVRAS-CHAVE: *Alucinação. Conteúdo perceptivo. Disjuntivismo. Michael Tye. Mark Johnston.*

INTRODUCTION

It is widely accepted nowadays that, when we have a veridical perception, like seeing an apple before us, we are directly in contact with a particular object in the external world. The contact is *direct* because it is not mediated by any internal (or mental) item. By perceiving the apple before us, we are directly aware

[1] Departamento de Tecnologia em Computação e Humanidades (DTECH) e Programa de Pós-Graduação em Filosofia (PPGFIL) da Universidade Federal de São João del-Rei (UFSJ).

of it. Our awareness, as many like to put it, does not stop somewhere short of the external object. This is the view of *direct realism*.

Hallucination poses a puzzle to direct realism. When hallucinating, no external object is out there for us to be aware of. Nonetheless, we can still have a vivid experience that can be exactly alike, in every single detail, to a veridical perception. Traditional versions of the well-known *argument from hallucination* take the bold step from subjective indistinguishability to the conclusion that direct realism is simply false.

In response to this challenge, direct realists have often defended some sort of *disjunctivist* approach, according to which veridical perception and hallucination are radically different types of things and share no common mental state, even though they can be indistinguishable *from the perspective of the subject*. On the other side of the fence, the various versions of the argument from hallucination have usually inspired *conjunctivist* views, in which a common shared element explains subjective indistinguishability.

The obvious question that arises is, if anything, what a hallucination is an experience of. In other words, more akin to the currently prevalent representationalist parlance: what is, after all, the content of hallucination? This question presupposes that hallucinations have content. The very idea that experiences have content is a controversial claim.[2] For present purposes, however, I simply take it for granted. I assume here that experiences have representational content of some sort. I take this bold claim for granted, as my starting point. One must start somewhere, and this is as good a start as it can be, given where I want to get to.

Amidst his investigations on perceptual content, Michael Tye (2009), in one of his moods, proposed an account of the content of hallucination.[3] Somewhat independently, Mark Johnston's (2004) analysis of the object of hallucination also provides us with elements to think about the content of hallucination (or so I shall argue). As a matter of fact, Johnston (2014) explicitly rejects the view that experiences have content. Though he is not himself a representationalist, I follow Hilbert (2004) in considering that his view on hallucination is interestingly adaptable to the representationalist framework. Given that the debate concerning the content view is entirely beside my point here, I opted to speak as if Johnston were happy with the content talk, which isn´t in fact the case. The use of a common representationalist parlance avoids unnecessary complications and eases the way to the relevant goals here.[4]

---

[2] This claim is famously denied by Travis (2004), Martin (2004, 2006), Brewer (2004, 2006, 2011), Johnston (2014), among many others.

[3] Anyone familiar with Tye's work knows that his views have changed significantly throughout the years. This is no different when it comes to hallucination. I consider here the account that Michael Tye advanced in the 2008-2009 period. After that, he (2014) published a paper proposing a quite different account of hallucination. Given my scope and interests here, I won't take this latest paper into account. Whenever I refer to Tye's view, I mean his 2008-2009 account. The novelties that came later, in any case, do not touch the points discussed here.

[4] I thank the second anonymous referee for pressing me to make this point even more clear and explicit.

Tye's and Johnston's perspectives are alike in many respects, but there are important distinctions that may lead to different accounts of the content of hallucination. This paper is structured in the following way: (I) I start with some preliminary distinctions; (II) Tye's account is briefly introduced; (III) Johnston's view is presented; (IV) with the stage already set up, I draw some comparisons and evaluate their views; (V) a radically different alternative (the epistemic conception of hallucination) is then considered and rejected; (VI) lastly, a sweeping metaphysical argument is considered and left open as a living challenge. At the end, if any illumination on the so far obscure content of hallucination can be reached, my aims have been achieved.

1 Preliminary distinctions

Both Tye (2009) and Johnston (2004) acknowledge that there is an important core of truth in direct realism. The view that Tye (2009, p. 541) prefers to call "naïve realism" is taken as a reasonable starting point for his investigations; and Johnston (2004) is explicitly carving out a direct realist approach that falls somewhere between disjunctivism and conjunctivism, as these terms are commonly conceived, and that can satisfactorily react to the argument from hallucination. In these very broad terms, the parallels between their views are quite evident.

In order to avoid terminological confusion, I initially unpack some of the key ideas loosely adumbrated so far. Since the discussion at stake walks along a fine line separating positions that are allegedly opposed, the conceptual tools must be precise enough to cut fine differentiations that are often neglected. To that end, I distinguish strong and weak versions of each of the main concepts being used here. A lot of misunderstanding seems to spring from the failure to acknowledge these distinctions.

The *strong argument from hallucination* is an attempt to refute direct realism. Here is one way of unpacking it:[5]

(1) Hallucination and veridical perception can be subjectively indistinguishable.

(2) If they can be subjectively indistinguishable, there must be some common element that explains it.

(3) The common element is some sort of mental state (or act of awareness).

(4) Hallucinations are mere mental states (i.e. they are not related to external objects).

(C) Therefore, the only act of awareness shared by veridical perception and hallucination must be purely mental (i.e. it cannot be a relation to external objects, nor can it include elements that are not mental).

[5] This version of the argument is based on the one presented by Aranyosi (2010).

The *weak argument from hallucination*, on the other hand, is restricted to the two initial steps of the strong argument. It only demands *some* common factor to explain subjective indistinguishability, but it is neutral on the nature of that common element. As far as it goes, the weak argument is consistent with direct realism. The strong argument, on the contrary, implies the denial of direct realism because the elements responsible for accounting for the phenomenal character of experience (both veridical and non-veridical) cannot include the direct or unmediated relation with external (or non-mental) items.

Strong conjunctivism is the claim that (i) there is a common element between veridical perception and hallucination that explains their subjective indistinguishability, *and* (ii) this shared element is an act of awareness. Strong conjunctivists typically distinguish between direct and indirect objects of awareness: the direct object is the one shared between veridical perception and hallucination, whereas the indirect object is the stuff in the external world that causes the veridical experience. This view is a straightforward consequence of the strong argument from hallucination.

Weak conjunctivism amounts to the milder claim that there is some common element between veridical perception and hallucination that explains their subjective indistinguishability. It is the counterpart of the weak argument from hallucination.

Strong disjunctivism consists in the negation of weak conjunctivism. It denies that there is *any* common element between veridical perception and hallucination that explains their subjective indistinguishability. Therefore, strong disjunctivists do not accept any version of the argument from hallucination, since the second step of the argument (shared by both weak and strong versions) is straightforwardly rejected. The natural upshot of this view is that veridical perception and hallucination have radically distinct contents (or, alternatively, that hallucinations have no content at all).

Weak disjunctivism rejects the claim that there is any act of awareness in common between veridical perception and hallucination. It rejects the stronger part of strong conjunctivism (i.e. the claim that there are direct and indirect objects of awareness). Weak disjunctivists only deny the strong version of the argument from hallucination. Consequently, this view is, in principle, compatible with direct realism and the weak argument from hallucination.

2 Tye's account of hallucination

Along his investigations, Michael Tye (2009) evaluates different accounts of the content of visual perception.[6] A significant part of his inquiry concerns whether or not particular objects should enter into the content.

[6] The focus on *visual* perception is harmless here. Though I confine myself to visual examples, the relevant point is general enough to be extended to any sensory modality.

According to the *existential thesis*, experiences have no singular content.[7] In this view, perceptual content is purely existential or general: it represents the world as having something or other with certain properties (e.g. size, shape, color) at some spatiotemporal location.[8] This view is mainly motivated by the requirement that perceptual content shall fix the phenomenology of experience. Since two experiences can be phenomenally alike (and, therefore, subjectively indistinguishable) even when their objects are numerically distinct (or when one of them has no object at all, as in the hallucinatory case), then the strictly *perceptual* content shall not be sensitive to phenomenally idle items such as particular objects.

The main argument advanced by Tye against this view presses its incapacity to accommodate the other constraint on perceptual content; namely, it fails to capture the accuracy conditions of experience. Tye (2009, p. 544) adapts an example from Grice (1961) to illustrate this point: a perceiver looks straight ahead and, unbeknownst to her, there is a mirror placed in front of her, inclined somehow so as to reflect a white cube that is out of her visual field. Now suppose that behind the mirror there is a red cube. Also unbeknownst to the perceiver, some special lighting conditions make the reflected white cube look red to her. This scenario leads the existential thesis to the wrong verdict that the experience is accurate, since the representational content is that there is something cubical and red at a certain location, and in fact there is. However, the red cube is obviously not seen, and the cube that is actually seen lacks the property of being red and at the perceived location. This is a clear case of illusion (or misrepresentation), and the existential thesis lacks the appropriate conceptual resources to explain that.

The defender of the existential thesis might attempt to avoid this problem, as Searle (1983) did, by adding the causal relation with the perceived object into the content. However, since the object itself does not enter into the content, this reply cannot avoid possibly deviant causal chains. Given the impossibility of capturing a particular item as such by descriptive means, the accuracy conditions cannot be fully captured in purely existential terms.

It can also be argued that Tye's argument begs the question against the existential thesis, since its defender can respond that the experience of the red cube in the example above is actually accurate: it says that there is a red cube in front of the subject, and that is actually the case.[9] It may be argued that the experience counts as an illusion not because of its content being falsidical, but because the object that is actually perceived is not red and in front of the subject. The perceived object is not itself determined by the content, so the illusory character of the experience has to do with the objects related to the experience, not with its content. However, the whole idea of perceptual content is to capture the satisfactory conditions of experience. How (if at all) these conditions will inform phenomenology is a separate issue. If there is something illusory about the

[7] This thesis is defended by McGinn (1982) and Davies (1992), among others.
[8] I assume here that the content is partly singular, since it includes particular places and times. This point is pressed by Tye (2009, p. 556) against the existential thesis. I make this assumption because otherwise the purely existential content would face obvious counter-examples.
[9] I thank the second anonymous referee for raising this point and pressing me to address this alternative response.

experience above, and its content fails to capture it, then there is something missing in the content. The missing part may be the perceived object, or the appropriate relation between subject and the perceived object. Insofar as content goes, it must explain perceptual error, whether or not the posited elements have any phenomenal impact. The problem with the existential thesis, as pointed out by Tye, is that it fails to deliver the right satisfaction conditions.

Once the existential thesis is rejected, Tye (2009) leans towards a position that has been traditionally taken by direct realists.[10] According to that position, the only way to account for our direct contact with external objects is to include them as components of the perceptual content. In the case of veridical perception, this strategy is quite compelling. However, it becomes much less appealing when it comes to hallucinatory experiences.

In order to have a perceptual content that accommodates both veridical perception and hallucination, Tye (2009) advances what he calls the "singular (when filled) thesis" (SWF henceforth).[11] This thesis claims that, in a veridical perception, the perceived particular object enters into the content, whereas in a hallucination, the content is just like in the veridical case, except that instead of a particular object there is a gap, or an empty slot, in the content. The contents of both veridical perception and hallucination share a common structure, that Tye (2009, p. 546) calls a "content schema". The SWF thesis captures the adequacy conditions in the following way: in a veridical perception, a particular object is represented as having some properties, and it in fact has these properties. The perception is illusory if the object lacks some of the properties attributed in the content. In that case, the experience is falsidical. In a hallucination, there is no particular object to fill out the content, and the gappy content can be understood as immediately falsidical, no matter which properties are represented.[12]

Tye (2009, p. 553) also motivates his adoption of the SWF thesis by the fact that the gappy content provides a more intuitive explanation of the "deceptive nature of hallucination". In a hallucination, the perceiver can be completely deluded, so as to take the hallucinatory object for a real one and react accordingly (e.g. the subject who jumps back to avoid a hallucinated spider). A deceptive hallucination is not perceived as a bunch of qualities, but as a real thing out there. Even though Tye initially qualifies this argument as "unpersuasive", he (2009, p. 553) affirms that "the supposition that there is gappy content in hallucinatory cases preserves as much similarity as can be preserved between those cases and the veridical ones". He even adds later that, because of its deceptive quality, "it seems that we really must suppose that his [the subject of a hallucination] experience has a gappy content, one with a **quasi-**singular character" (TYE, 2009, p. 555, author's emphasis).

[10] McDowell (1994), to name only one, is a tenacious advocate of that position.
[11] Similar versions of this thesis were defended by Burge (1991), Loar (2003) and Bach (2007).
[12] To be more precise, the gappy content can be considered falsidical or neither true nor false. For brevity, I consider it here as immediately falsidical. I also ignore for the moment complications such as the cases of *de re* hallucinations, in which a particular object is identified as the hallucinated object (e.g. a subject who hallucinates his mom entering in the room). I believe, as does Tye (2009, p. 548), that such cases pose no special difficulty for the thesis under discussion.

Now take the mirror example again. The SWF thesis has a straightforward explanation of the illusory nature of that experience: since the perceived particular object enters in the content, the perceptual content includes the white cube in it, not the red one. However, the illusory experience, in this case, seems to be *accidentally veridical*, since the experience says that there is a red cube in front of the perceiver, and in fact there is a red cube there. The problem is that the SWF thesis classifies the experience as illusory and gives the *unequivocal* verdict that the experience is simply falsidical or inaccurate, even though the world seems to be just as it is represented to be. The same problem arises in cases of veridical hallucination, in which the perceiver hallucinates something (a red cube, say, in front of her), and it turns out that there is a real red cube in the exact same location. As it stands, veridical illusions and hallucinations seem to be unsatisfactorily explained by the SWF thesis.

Let us look more closely at the veridical hallucination case. Tye (2009, p. 557) claims that, in this case, the gappy content disposes the perceiver to form an accidently true belief: "cases of veridical hallucination are veridical, then, only to the following extent: the visual experiences they involve dispose their subjects to form true beliefs. The experiences, however, are falsidical or at least neither true nor false". Tye adopts the view that "the relevant contents, thus, are potential cognitive contents and not actual visual contents of my experience" (TYE, 2009, p. 558, author's emphasis). Consequently, in a veridical hallucination, the perceiver, based on her perception, forms a higher-order cognitive state (a belief, say) that is accidently veridical. The perceptual content, however, is itself falsidical.

To sum up, Tye (2009) claims that the difference between a veridical experience and a hallucination is the following: in the first case, there is a particular content, in the second, a gappy one. The gappy content may dispose the subject to believe that there is a real object out there. Both particular and gappy contents share the same "content schema". Besides sharing the same content-structure, they share, if they are subjectively indistinguishable, the same "non-object-involving properties".

The resulting view asks for a proper understanding of the representationalist thesis. Tye (2008, 2009) draws a distinction between weak and strong versions of representationalism. Strong representationalism identifies phenomenal character and representational content, whereas weak representationalism claims that the phenomenal character only supervenes on the representational content. Once the subjective indistinguishability between veridical perception and hallucination is spelled out in terms of a shared phenomenal character, it follows that the SWF thesis is committed to denying the strong version of representationalism. This is so because, in this view, different contents can be attributed to phenomenally identical experiences.[13] Tye is well aware of this consequence, and is happy to

[13] Another option to combine the SWF thesis and strong representationalism would be to deny that the veridical and hallucinatory experiences are phenomenally identical. I thank the second anonymous referee for raising this point, but I prefer to consider this option latter, when I discuss the epistemic conception of hallucination. As of this moment, I prefer to follow Tye's steps and presuppose the possibility of veridical and non-veridical experiences being phenomenally identical.

embrace it. He (2009) holds that the phenomenal character of experience is given by the "cluster of properties" that is possibly shared by veridical and hallucinatory experiences. Although singular and gappy contents are quite different, they may share a cluster of properties that, ultimately, explains the phenomenological sameness of the two experiences. Weak conjunctivism, therefore, is vindicated: there is something in common, and this thing is not a mental state (or an act of awareness). In conformity with that, Tye (2009, p. 562) claims that "the solution is to look at the properties represented to find phenomenal character, and not to the representing of those properties". Given that the existence of a common mental state is denied, Tye (2009, p. 562) qualifies his SWF thesis as "a form of disjunctivism". In my terms, his position qualifies as weak conjunctivist and weak disjunctivist, and the strong versions of both theses are denied.

3 Johnston's account of hallucination

In many aspects, the differences between Michael Tye's (2009) and Mark Johnston's (2004) accounts of hallucinatory experiences are a matter of detail. The deeper dissimilarities only emerge after some effort, and also due to some liberty on my part to extend Johnston's thought *contra* Johnston and beyond the limits of his own investigations.

Johnston (2004) explicitly vindicates the weak version of the argument from hallucination, and he consistently considers it in its weak sense. Motivated by this argument, Johnston (2004, p. 114) demands a common factor between veridical perception and hallucination in order to account for the "(i) subjectively seamless transitions between certain cases of sensing and hallucination, and (ii) the distinctive character of hallucination itself".[14]

The term 'conjunctivism' is used by Johnston (2004, p. 114) in the strong sense, as meaning that (i) there is a common object of awareness between hallucination and veridical perception, *and* (ii) in the veridical case, the relation between external object and act of awareness is a causal one.[15] Noticeably, the distinction between mental item (which is the direct object of experience) and external item (which is the indirect object of veridical perception) is built into the very definition of the conjunctivist view. When Johnston rejects the conjunctivist approach, he is therefore rejecting *strong* conjunctivism.

Strong conjunctivism, as I use this term here, contradicts direct realism. When criticizing this view, Johnston (2004, p. 119) claims that "when we see, or

---

My aim right here is less to explore all the representationalist options and more to characterize Tye's stance.

[14] By "subjectively seamless transitions", Johnston (2004) means the fact that veridical and hallucinatory experiences do not carry with them, necessarily, any distinctive phenomenological mark that could be used by the subject of the experience to tell them apart. In other words, a subject could go from a hallucination to a veridical perception, and vice-versa, without possibly being able to tell the difference between them by means of what is phenomenally given by the experience itself. [I thank the first anonymous referee for pressing me to clarify this point.]

[15] Hilbert (2004, p. 187) criticizes Johnston's inclusion of the causal connection in the very definition of the conjunctivist view, and I am inclined to agree with his criticism. This point, however, goes much beyond my scope here.

more generally sense, external particulars those particulars are no less 'directly' present to us than anything in hallucination". The reason why Johnston rejects strong conjunctivism is somewhat like the reason that led Tye (2009) to defend the singular content of perception.[16] According to Johnston (2004, p. 119), "without external particulars immediately seen or sensed, the whole scheme of descriptive identification of particulars would be ungrounded". The crucial point here, shared by both philosophers, is that particulars are not derivative objects of awareness.

Johnston (2004, p. 121) initially characterizes disjunctivism in its weak sense, as the claim that there is no common act of awareness between veridical perception and hallucination. However, he later switches to the strong version, claiming that, in the disjunctivist view, there is no common factor explaining the subjective indistinguishability between veridical and hallucinatory experiences. When he rejects disjunctivism, he has in mind the strong version. He justifies this rejection in the following terms:

> The crucial point is that the something in common need not be an **act of awareness** of which seeing is a subspecies. There can be a common **element** in awareness, which explain seamless transitions and so forth, but which is not itself a common **act** of awareness. (JOHNSTON, 2004, p. 170, author's emphasis).

Enough for terminological clarifications. We can safely conclude by now that both Johnston and Tye only reject the strong versions of conjunctivism and disjunctivism.

Johnston's (2004) investigations are mainly guided by the following questions: what explains the seamless transition (from the subject's perspective) from a case of veridical perception to a case of hallucination? And what kind of thing, if any, is a hallucination related to?

These questions are interwoven. If hallucinations are related to anything at all, that *relatum* is supposed to fix the phenomenology of experience and account for the seamless transitions from veridical to hallucinatory experiences. Disjunctivists, by contrast, typically claim there is nothing properly perceptual that could count as the *relatum* of a seemingly existing hallucinatory object. Some of them resort, for that matter, to higher-order cognitive states in order to explain the nature of hallucination.[17] For those resorting to this strategy, a hallucination is akin to a false belief about what is seen, or a case of *seeming to be seeing* something. Disjunctivists of this sort, such as Huemer (2001, p. 127), claim that "hallucination

[16] I was convinced by the second anonymous referee to weaken my former claim that their reasons for rejecting strong conjunctivism are "very much the same". They are alike insofar as both claim that the perceived object is a constitutive part of veridical experience, and that rules out the possibility of characterizing veridical and hallucinatory experiences in all the same terms. However, as pointed out by the very attentive and helpful referee, whereas Tye is more directly concerned with the correct metaphysical account of the perceptual content, Johnston is more directly motivated by epistemological considerations concerning what we can learn from different kinds of experiences. The underlying epistemological motivations are made more explicit in Johnston (2006).

[17] This position will be considered more closely in section V.

is not awareness at all […] for awareness is a relation between the subject and the world, and the hallucination fails to have the right relational properties".

Another disjunctivist line also denies the existence of any *relatum* for hallucination, but appeals to merely intentional objects. In the same way that Ponce de Leon's search for the Fountain of Youth does not demand any existing object for him to be searching for, hallucinations may relate us to merely intentional objects. However, Johnston (2004) claims that this analogy is infelicitous. Hallucinations, like veridical perceptions, have a sensory nature and present items to the subject's attention. In the case of 'searching', on the other hand, there is only a verb that takes a grammatical object. One can obviously search for something that does not exist. In the hallucinatory case, however, the urge to determine an object (or *relatum*) does not come from the need of finding a grammatical object to the verb 'to hallucinate'.

The reason why Johnston (2004, p. 129) advocates an "act/object analysis" of hallucination is because hallucinations "serve up distinctive items for demonstration", and from these items, he claims, "we can learn certain novel things". The question is how this object (or *relatum*) of hallucination shall be conceived. The following considerations guide his enterprise: (i) hallucinations are not original sources of *de re* thoughts about particular objects; (ii) hallucinations can secure original reference to qualities, so they can ground *de re* knowledge of qualities; (iii) no particular object can be the primary object of hallucination, but, in a certain sense, particular objects can be considered the secondary objects of hallucination.

Unless one's ontology is open to accept non-existing entities and/or sense-data, the first consideration shall be uncontroversial. Hallucination is not a relation with any particular object, and that is why it is so puzzling to direct realists. Since there is no particular object of hallucination, there can be no *re de* thought about particulars grounded on hallucinations.

Commenting on the second consideration, Johnston (2004, p. 130) claims that "Frank Jackson's Mary could come to know what red is like by hallucinating a red thing or by having a red afterimage". If qualities are *directly* presented to the hallucinator, as Johnston claims to be the case, then *de re* knowledge of qualities can be grounded on hallucination. This claim is, indeed, his main motivation for adopting an act/object analysis of hallucination. If some kind of *de re* knowledge can find its ground on hallucinations, there must be a *res* to which a hallucination is a relation to. There must be, in this sense, an object of hallucination that constitutes its content.

A powerful argument in favor of the claim that a subject can hallucinate novel qualities (and then, based on the hallucinatory experience, learn how these qualities look like) is the following experiment. After being exposed to a bright monochromatic unique green light in a dark room for about twenty minutes, the room is illuminated and the subject afterimages a small red patch, which is then superimposed on a small red background, causing the subject to have a

supersaturated red afterimage.[18] The supersaturated red is more saturated than any visible red in normal circumstances. This is a color that can never be seen, but only afterimaged. This experiment confirms the thesis that novel qualities can be assessed by experiences involving no particular objects instantiating the represented properties (such as hallucinations and afterimages).

The third consideration, that distinguishes primary and secondary objects of hallucination, elaborates on cases of alleged *de re* hallucinations of particular objects. The primary object is the cluster of properties hallucinated. The secondary object, which would account for the *de re* nature of the hallucination, includes references to particular objects. Johnston (2004, p. 132) claims that, in such cases, the primary object simply "strikes the subject" as being about a certain particular object. Particularity here, however, is merely derivative, being based on the "subject's existing repertoire of singular reference" (JOHNSTON, 2004, p. 132). Secondary objects of hallucination are, in fact, just a "*façon de parler*" (JOHNSTON, 2004, p. 143). The only genuine objects of hallucination are the primary ones. They are, strictly speaking, not objects, but clusters of properties.

Taking into account the considerations above, that aim to uncover the seemingly obscure nature of hallucinatory experience, Johnston (2004) develops a theory that attempts to explain, among other things, the subjective indistinguishability between veridical perception and hallucination. In a veridical perception, the sensed scene before the eyes has a certain relational and qualitative structure that is instantiated by particular objects. The scene itself can be understood as a scene type, which he (2004, p. 133) calls a "sensible profile". The sensible profile is a complex of qualitative and relational properties that explain the way the scene looks to the perceiver. This *way the scene looks* involves a certain *layout*: "whichever particulars are implicated they have to stand at certain times in certain positions in a three-dimensional space at certain directions and distances from your position now" (JOHNSTON, 2004, p. 134). Although the layout includes particular places, times, and subjects, it is understood as purely relational in itself, as "a universal rather than a particular" (JOHNSTON, 2004, p. 134). Different things can instantiate the same layout. In veridical perceptions, the sensible profile involves more than the layout: it also includes particular objects that fill out the layout.[19]

> Your seeing the scene before your eyes is your being visually aware of a host of spatio-temporal particulars instantiating parts of such a profile or complex of sensible properties and relations. The suggestion is that in the corresponding case of a subjectively indistinguishable hallucination you are simply aware of the partly qualitative, partly relational profile. This means that the objects of hallucination and the objects of seeing are in a certain way akin; the first are complexes of sensible qualities and relations while the second are spatio-temporal particulars instantiating such complexes. (JOHNSTON, 2004, p. 135).

[18] The experiment is explained in more detail in Johnston (2004, pp. 141-2), who took it from Hurvich (1982, pp. 187-8).
[19] Johnston (2004) also includes natural kinds in the content of veridical perception. Nothing in this paper hinges on that, and I prefer to set it aside and remain neutral on that.

The object of hallucination is a "proper part of the more demanding sensible profile that one is aware of in a corresponding case of seeing" (JOHNSTON, 2004, p. 136). In both cases, the subject is *directly* aware of something. The difference is that in a hallucination one is directly aware of less than one would be aware of in the corresponding veridical perception.

Hence, subjective indistinguishability is explained by a common factor that is not a common act of awareness. Johnston's strategy parallels Tye's (2009) in many respects. However, Johnston still has to explain what Tye (2009, p. 553) calls the "deceptive nature of hallucination". In fact, the perceiver can be completely deluded by a hallucination and react as if there were a particular external object being perceived. Hallucination may well be incapable of securing *de re* reference to particular objects, but it is not perceived as a bunch of floating qualities. Johnston (2004, p. 140) is perfectly aware of this demand: "hallucinated sensible profiles can mimic particularity". This mimicking capacity is explained by the spatiotemporal layout generating the illusion of a particular moving in certain directions. The perceiver is led to believe that there is a particular object out there, but this seeming object is a secondary object of hallucination, an object that appears only in higher-order states. The content of hallucination itself contains only a complex of sensible qualities and relations. According to Johnston (2004, p. 142), "thanks to containing certain properties in certain relations to continuous places and times, a primary object can immediately strike the subject as a moving particular". As a result, the deceptive nature of hallucination is something like a Vegas billboard illusion: it looks as if an object is moving around the board, when in fact there are only successive lights going on.

4 Taking stock

The structural similarities between the views of Tye (2009) and Johnston (2004) are striking and, in fact, it is sometimes a tricky job to say exactly what difference there is, if any.

The first likely dissimilarity concerns the gappy content theory. The Vegas billboard picture evoked by Johnston's explanation of the deceptive nature of hallucination is compatible with an existential account of the content of hallucination (though not a *purely* existential one, since the layout includes particular times and places). The idea that hallucination "mimics particularity" seems, in principle, perfectly compatible with an existential account. In fact, this account even seems to provide a theoretical gain in metaphysical economy, since the existential approach of the hallucinatory content does not ask for any (possibly costly and complicated) "metaphysics of empty slots".[20] Assuming that veridical perception is constituted by singular objects, which is granted by both Tye and Johnston, the existential account of hallucination can find no place for the notion

[20] Tye (2009, p. 548) does not elaborate on this topic, but he recognizes that this is a real issue and must be eventually addressed.

of a common schema (the "SWF schema") shared by both veridical and hallucinatory experiences. It is hard to see, though, what sort of explanatory role this notion is actually playing. Since the common factor role, which explains subjective indistinguishability, is assigned to non-particular elements in the content, it is far from clear what reasons there are, if any, for insisting on a common (singular-like) schema.

Still, the gappy theory seems to provide a better account of veridical hallucination. In the case of veridical hallucination, Tye (2009) argues that the precisely *perceptual* content of the experience is falsidical, though it disposes the perceiver (at least in so far as perception goes) to form an accidently veridical belief. In contrast, according to the existential account of hallucination, the strictly perceptual content of a veridical hallucination turns out to be simply veridical. In this view, there is nothing in the perceptual content of hallucinations showing that something went wrong in perception. However, a hallucination (be it accidently veridical or not) obviously involves some sort of defective encounter with reality. If hallucinatory content is cashed out in terms of (instantiated) clusters of properties spatiotemporally located, then we lack the required resources to explain what went wrong in veridical hallucination.

Another argument against the existential account of hallucination comes from phenomenological considerations. After analyzing the phenomenological elements that determine the "sense of reality" of perceptual experience, Dorsch (2010) noticed that some elements can hold in the absence of others. Among the "reality characteristics" distinctive of veridical perceptions and of (seemingly veridical) hallucinations, in opposition, say, to typical imagining or dreaming, there are two of major interest for us: (i) *particularity* (objects are experienced as being numerically distinct), and (ii) *locatedness* (perceived objects appear to be spatiotemporally situated). Those two features, however, do not go necessarily together. Some cases of seeing, for instance, are vague about the precise location of the object. A limiting case is recounted by Sims (1995, p. 110): a patient with histrionic personality disorder[21] vividly hallucinated a person at her bed, but she was unable to locate that person spatially, in relation to her environment. When asked to do so, she said she couldn't, since the hallucinated person had no definite location in relation to the other objects in the room (walls, curtains). The case recounted by Sims is somewhat anecdotal, and the whole situation seems to be quite underdescribed. Though this example may be less than persuasive, it seems less unlikely that other cases like that may exist, which casts doubt on Johnston's (2004) attempt to explain the feeling of particularity in hallucinations as derived from the feeling of locatedness.

To be fair, the existential account of hallucination is only one possible way of elaborating on Johnston's ideas. He is, to be true, against the idea of experiences having content altogether. Still, a Johnston-inspired representationalist view seems to fare better with gappy contents for hallucinations. If we plug the gappy account

[21] Histrionic Personality Disorder (HPD) is defined by the American Psychiatric Association (2013, p. 667) as a personality disorder characterized by a pattern of excessive attention-seeking, emotional overreaction, and over-dramatization of ordinary situations.

in Johnston's theory, we can simply regard the uninstantiated sensible profiles as analogs of gappy contents. In this (gappy-representationalist) reading of Johnston, layouts demarcate spatiotemporal gaps that are mapped into the content.

As to the phenomenological objection mentioned above, it may be argued that locatedness suffices to fix numerical identity and to enable demonstrative reference. In that case, locatedness would be sufficient to determine the sense of particularity, although it may not be necessary for it (the feeling of particularity could well have other sources). Another possible reply could appeal to the fact that vague locations are still locations, and a certain degree of locatedness would be enough to generate the sense of particularity. In any case, by adopting the gappy approach, whatever vantage point this may offer concerning the explanation of the phenomenology of particularity, Johnston can just as well claim the same.

5 The epistemic conception of hallucination

Up to this point, we have just assumed that veridical and hallucinatory experiences can share the very same phenomenology. However, the fact that two experiences cannot be told apart introspectively is consistent with their phenomenal character being quite different. This possibility is curiously reinforced by the non-transitivity of indistinguishability, which is remarked by Johnston (2004, p. 165). If someone hallucinates a dark red patch that becomes gradually less saturated, one may be unaware of the difference between two patches presented in brief successive instants. Nonetheless, the initial and final moments present patches that are clearly distinguishable. Johnston (2004, p. 166) notes that "the hallucinator can miss some of the qualitative features of his hallucination", since there can be "more to the object of hallucination than how it strikes the subject". This observation is used to support his act/object analysis of hallucination, but it can just as well be used to motivate the dissociation of what the subject can introspectively differentiate from real differences in phenomenal character. Hilbert (2004, p. 188), for instance, takes the failure of transitivity in perception to motivate the denial of the naïve claim that "the immediate objects of perception are just as we taken them to be".

The possibility of veridical and non-veridical experiences having different phenomenal properties, despite their being subjectively indistinguishable, was largely explored by the so-called *epistemic conception of hallucination*.[22] This view proposes a more radical disjunctivist account of perceptual experience in which the explanation of subjective indistinguishability does not resort to a common phenomenology, but to a certain epistemic process of introspection in which different things may simply look alike.

There are, certainly, many players in this game. I have no intention to exhaust the alternatives here. In what follows, I illustrate this position considering two possible routes. The first way, which is the most influential one, characterizes

[22] This view is defended, among others, by Martin (2004, 2006), Soteriou (2005), Brewer (2008), and Fish (2008, 2009).

hallucination "solely by saying that it is like what it is not" (DANCY, 1995, p. 436). This is the *negative epistemic conception of hallucination*. The second way is even more radical and claims that hallucinations are false beliefs about experiences, but have no phenomenology of their own. This is the *eliminativist epistemic account of hallucination*.

The most prominent proponent of the negative approach is Michael Martin (2002, 2004, 2006). He claims that a hallucination consists fundamentally in an experience that is subjectively indistinguishable from a veridical experience from the perspective of the perceiver. A hallucination is, therefore, something that fundamentally looks like something it is not, and its nature consists uniquely in being a kind of impostor. Hallucination, in this view, is defined in terms of subjective indiscriminability, which is in turn characterized in terms of knowability. An experience is subjectively indiscriminable from a veridical experience of a certain kind if and only if it is not possible for the subject to *know by introspection alone* that her experience is not of that kind.[23] According to this view, phenomenal sameness and subjective indiscriminability are distinct but closely related phenomena. If two experiences are phenomenally identical, it follows that a subject with well-functioning discriminatory abilities will not be able to tell the difference between them. But the converse does not hold. Two experiences can be subjectively indiscriminable to a subject even if they are not phenomenally identical. This is so, for example, if the phenomenal difference is too slight and therefore inaccessible to the subject, even if her introspective abilities are functioning properly.

A theory of hallucination must offer the conditions that must be satisfied in order for a state to count as hallucinatory. Siegel (2004) pointed out that there are obvious counterexamples to the negative epistemic definition. Consider the case of cognitively unsophisticated hallucinators. A toad, for example, may not be able to know anything at all by introspection alone, since introspection involves higher-order representations that cognitively simple creatures like toads may not be capable of. In this case, a toad trivially satisfies the condition above: it is never possible for the toad to know by introspection alone that its experience is not a veridical one. As a consequence, toads (and rocks and tables) would be trivially hallucinating all sorts of things all the time. This is obviously absurd.

Martin (2006, p. 379) responded to this objection by cashing out subjective indiscriminability in terms of *impersonal* knowledge. The idea is to replace the particular subject with an *ideal* introspector. The improved formulation of the conditions is the following: an experience is subjectively indistinguishable from a veridical experience of a certain kind if and only if it is not possible for an *ideal introspector* to know by introspection alone that the experience is not of that kind.

Nonetheless, the "impersonal" version brings other difficulties with it. Brewer (2011, p. 111) compared the idealized notion of "being indistinguishable by introspection" with the mathematical notion of "being unknowable". This

[23] The term 'introspection' denotes here the distinctive way in which the subject comes to know about her own mental states, whatever that ability exactly consists in.

comparison, however, is problematic. Pautz (2010, p. 275) pointed out that there is an important difference between these cases. In the mathematical context, "being unknowable" is intuitively grasped within the idealized mathematical framework. In the context of perceptual experience, however, we have no basic pretheoretical intuition to appeal to. In the context of perceptual experience, the impersonal condition ("being indistinguishable by introspection") is not an epistemic condition at all, but it is an entirely primitive notion. As a primitive notion, it is not defining hallucination in terms of something else that we have an independent grasp on. Hallucinations, on this view, become some sort of brute facts that resist any deeper explanation.

There is also another reason why the idealized condition seems problematic. Being subjectively indistinguishable is accounted for in terms of what an ideal distinguisher can distinguish. This seems circular. This would only be informative if the kind of ability of the ideal distinguisher were defined in independent terms. This is supposedly done by the notion of "knowledge by introspection". But this notion, as pointed out above, is unclear and cannot be explained in independent epistemic terms. As a primitive notion, it only renames the concept that is supposedly being defined. The property of being subjectively indistinguishable is defined in terms of the property of being unknowable by introspection, but the *explananda* is as primitive and non-intuitive as the *explanandum*. This is why the definition seems circular: whatever you take to satisfy the first condition you can make it satisfy the second one, for there is no independent procedure that can be used to test if something satisfies only the second condition. The right side of the bicondicional does not explain anything: it is itself in need of explanation.

Responding to Siegel's (2004) criticism concerning the hallucinations of cognitively unsophisticated creatures, Martin (2006) says that a simple creature does not need any capacity to introspect: we can "attribute experience to the dog through attributing a specific take on the world, without thereby supposing that the dog is self-aware" (MARTIN, 2006, p. 396). His negative epistemic condition says that the dog has a hallucinatory experience if and only if an ideal introspector having the same experience would be unable to tell it apart from a veridical experience. The pressing question, however, is what distinguishes this (epistemic) condition from the property of being subjectively indiscriminable from some matching veridical experience.

Moreover, the negative epistemic criterion seems insufficient to come to grips with the nature of hallucination. Smith (2008) points out that many ordinary non-hallucinatory experiences can meet the negative criterion (i.e. they are subjectively indiscriminable from veridical experiences). The reason why the negative epistemic condition does not satisfactorily demarcate the class of states that deserve the label 'hallucination' is because it is inadequate to pick out exclusively sensorial states, and hence inadequate to distinguish hallucinations from non-sensorial states. Consider, for example, the experience of a very rapid flash of light.[24] The subject of such an experience can well wonder: 'did I just see

[24] The example is from Smith (2008, p. 184). This everyday case can also be found in psychological experiments using a tachistoscope, which is a device that flashes images on a screen very briefly.

a flash?' The situation here admits of three explanations. Maybe the subject did see a flash, but she is not sure because it was barely detectable. The experience was, so to say, at the very threshold of her discriminatory abilities. Another possibility is that the subject did not see anything, but was simply 'under the impression' that she did. A third possibility is that the subject briefly hallucinated a flash of light. The main difference between the second and the third cases is that in the third case the subject had a sensory experience, whereas in the second one she had no sensory experience at all, however brief. According to Smith (2008, p. 185), "there are here three psychological states that need to be distinguished from each other: having a momentary perception, having a momentary hallucination, and having neither, but merely 'thinking' that one has, or may have, just perceived something". The problem of a negative epistemic criterion is that it fails to distinguish hallucinatory experience from mere 'thinking'. Merely thinking that you have a sensorial experience does not amount to effectively having one, and hallucinating, contrary to mere thinking, is intrinsically sensorial.

The negative epistemic view is not only in trouble in finding a plausible criterion to demarcate hallucinatory experiences in a non-circular way. Martin (2002) defends the thesis that perceptual phenomenology extends beyond what is discriminable to the subject. According to him, the "phenomenal nature" of a perceptual experience outstrips what is subjectively distinguishable, or the "phenomenal character". In his terminology, even though veridical experience and hallucination may share the same "phenomenal character", they differ in "phenomenal nature". He accuses the representationalist of believing in the myth of a common nature. Two different things, with different natures, can surely be subjectively indistinguishable. The property of being subjectively indistinguishable from something else does not pick out a ground-floor psychological type. The only thing that unifies the class of states that are subjectively indistinguishable from veridical experiences is the very property of being subjectively indistinguishable. Martin's account, however, does not explain *why* hallucinations look so similar to veridical perceptions. The negative account that he advocates can (at most) tell which states count as indistinguishable, but the fact that those states are indistinguishable is simply a brute fact of the world. What seems particularly problematic about his view is that the phenomenal nature of hallucination is simply left aside, as if no positive account of it could possibly be given. Unless an account of the metaphysical ground of hallucination is given, the fact that hallucinations have the sensorial phenomenology that they have is simply left unexplained.

The difficulty to come to grips with the phenomenal nature of hallucination by means of a negative strategy has led some philosophers to try a more radical route: what if we simply deny hallucinatory phenomenology altogether? This gave birth to what I call the *eliminativist epistemic account of hallucination*.

William Fish (2004; 2008; 2009) is perhaps the most prominent defender of this view. He claims that subjective indistinguishability can be fully explained by the "discriminatory context", or "the subject's discriminatory capacities and the observation conditions under which the discrimination is attempted" (FISH, 2008,

p. 146).[25] Different things can be indistinguishable to a subject at a time. The subjective indistinguishability between hallucination and veridical perception, he claims, is just a topic for empirical investigation. One possible explanation is that it is generated by a deficit in the meta-cognitive skill of "reality discrimination".[26] According to Fish (2008, p. 157), "reality discrimination is characterized as the ability we have of telling mental episodes that are internally generated apart from real veridical experiences". By themselves, he claims, hallucinations have no phenomenal character. Hallucinators are mistakenly led to believe that they have visual experiences, with specific phenomenal characters, but "although such subjects think/believe/judge that hallucinatory states have phenomenal character, they are wrong" (FISH, 2008, p. 159).

Although hallucinations have no phenomenal character on their own, there is obviously something it is like to hallucinate. However, according to Fish (2008, p. 160), the explanation of *something-it-is-like claims* is simply that, when hallucinating, the subject falsely believes that she has a *perceptual* experience. Just like (visual) perceptual experiences prompt beliefs that "there is something it is like to see something", the phenomenal impression may just as well be the upshot of a false perceptual belief (FISH, 2008, p. 160). By so doing, Fish explicitly inverts the "standard order of explanation".

Fish (2008) goes even further. If the subject in question is not conceptually sophisticated enough to form beliefs (e.g. animals or infants), then the allegedly hallucinatory experience is nothing more than a behavioral reaction to some cognitive or perceptual malfunctioning.

To begin with, I must confess that the epistemic conception of hallucination strikes me as quite bewildering. The very possibility of inverting the "standard order of explanation" is hard to swallow. One must be very eager to vindicate strong disjunctivism to end up defending such an unlikely story. The epistemic conception is committed to the odd view that higher-order cognitive states can generate subjective states that are exactly like actual perceptions, but that lack phenomenal character altogether. Hallucinations are thus some sort of shadows from beliefs. Shadows that acquire a seemingly phenomenal quality just because the subject believes so. In what follows, I summarize a few arguments against this view.

The epistemic conception, at least in Fish's (2008) version, is unable to account for cognitively unsophisticated hallucinators. In Fish's view, unsophisticated creatures that lack higher-order states like beliefs (due to their lack of the appropriate concepts or mechanisms) would not be able to have hallucinatory experiences. Susanna Siegel (2008) argues that Fish's behaviorally-based (or "effect-based") explanation of animals' hallucination is deeply unconvincing. She (2008, p. 215) remarks that his theory "does not ensure that hallucinations have any felt reality from the point of view of the hallucinator". In the case of unsophisticated creatures, hallucination (if it still deserves this title at

[25] For the record, Fish's notion of 'indistinguishability' is strongly influenced by Williamson (1990).
[26] This term is taken from Slade and Bentall (1990, p. 125).

all) lacks not only a proper phenomenal character, but there is nothing it is like to be in that state. A lethargic cat that hallucinates a butterfly but remains quiet would be a theoretical impossibility in this view. For that matter, I quote Johnston (2004, p. 124):

> Being susceptible to visual hallucination is a liability which just comes with having a visual system, i.e., comes with being able to see, and does not require the operation of the ability to think or believe or reflectively grasp the fact that you are seeing, any more than seeing requires this.

Moreover, impossible scenes, such as Escher's drawings, can perfectly well be objects of hallucination (SIEGEL, 2004). In such cases, one could know, only by introspecting the scene, that it cannot be veridical. Consequently, it would be irrational to believe, based on introspection, that this is a veridical experience of any sort. Since the corresponding higher-order state cannot be made credible, it makes no sense to explain this hallucination as a projection from a belief that cannot be rationally believed. If I don't believe in what I see, I don't believe that I'm having a perceptual experience. Consequently, I should not believe there is something it is like to be in that state. This case seems to break the explanatory chain of the epistemic theory in its very origin.

The epistemic conception of hallucination, I conclude, does not seem promising. The negative and eliminativist strategies face pressing difficulties, and I fail to see how they could overcome them. Unless strong disjunctivists come up with a more persuasive account, we have good reasons to stick to weak conjunctivism, just like Tye and Johnston did.

6 Metaphysical worries

First of all, it shows the persistent difficulty in getting rid of the Cartesian dual world, divided between Johnston (2004) claims that the sensible profile shared by hallucination and veridical perception is a "proper part" of the content of veridical perception. The notion of being a *proper part* has strong metaphysical connotation. If the perceptual relation with the world, direct as it is, involves the representation of aspects of reality, then a proper part of it seems to be the representation of fewer aspects of reality. A hallucination (understood as an uninstantiated sensible profile) is not related to the world in a less direct way, but is only related to less.

Gappy and singular contents obviously differ from one another. Tye (2009, p. 562) claims that "at the level of content itself, there is indeed no common factor". But he remarks elsewhere that "the content involved in veridically experiencing a red object and the content involved in hallucinating a red object have something important in common" (TYE, 2008, p. 209). This *thing* in common brings with it a metaphysical difficulty. Johnston (2004) explicitly characterizes the common factor

as qualities and relations of sensible profiles, which are characterized as uninstantiated universals. When hallucinating, the subject is aware of universals. However, Dunn (2008, p. 378) remarks on how odd it is to be directly aware of universals, not to say *uninstantiated* universals. To start with, no causal relation can hold between (uninstantiated) universals and our awareness of them. According to Dunn (2008, p. 378), "it seems to be a process cloaked in mystery".

What, after all, accounts for the phenomenological presence of a certain quality in a hallucination? The representationalist answers this question by pointing to the content, and saying that the property is represented to be out there, but it happens not to be there. The representationalist can go even further and say that a given experience *e* represents a given quality *Q* in virtue of the fact that *e* is normally caused by training the eyes on real objects that instantiate *Q*. Indeed, it is because the subject undergoes *e* that she is ever enabled (if she is sophisticated enough) to ask 'what is that' with relation to *Q*.

Though ingenious, the representationalist approach seems to miss an important part of the whole story. Fish (2004, p. 8) subtly observed that a deeper question was left unanswered: after all, *why* does being in a state that represents a certain property suffice to make this property phenomenologically present? As he qualifies the question, it is not the "thin causal question" of "why the subject comes to be in state *S* (a question about the aetiology of the state)" (FISH, 2004, p. 9). We touch here what Fish calls the "deep explanatory question of **why state S has the phenomenology it does**" (FISH, 2004, p. 9, author's emphasis). The problem that must be addressed by any conjunctivist theory is how to "deep-explain" the phenomenological presence of properties without their actual instantiation in the perceived scene. Strong disjunctivists, like Fish (2004, p. 9), understand quality awareness as a real-world instance of the quality acting directly "on the subject's sense organs". Since they postulate a common factor to explain phenomenological similarity, conjunctivists, on the other hand, are committed to offering the same "deep explanation" for the phenomenological presence of a quality in both hallucination and veridical perception. As a consequence, in veridical perception, "the presence to the senses of certain visual properties – their perceptual presence – does not deep-explain why those properties are phenomenologically present" (FISH, 2004, p. 9).

The "deep problem" points to a profounder difficulty in combining conjunctivism and direct realism. Part of what we are directly related to in our acts of awareness are mundane properties. The argument from hallucination, traditionally applied to *objects* of experience, can be adapted and spelled out in terms of *properties*.[27] The argument, in rough, goes from the fact that one can hallucinate an uninstantiated property to the conclusion that the perceived property in veridical perception is not an instantiated physical property of external objects. Conjunctivists simply assume that hallucination and veridical perception are both related in the same way to the same kind of properties, but their assumption may be metaphysically unwarranted. Even if the properties represented

[27] That is exactly what Thompson (2008) did when arguing for the incompatibility of representationalism and direct realism.

in both cases are the same, there is still something mysterious about how they are related to the subject. Contrary to direct realism, it seems that the very instantiation of a property is irrelevant to its perceptual appearance.

Since hallucinatory experiences are not encounters with existing particular objects, representationalists must look elsewhere to find the grounds of their phenomenology. Given the principle of ontological parsimony (which, by the way, inspired the whole representationalist project of naturalizing consciousness), one shall refrain from populating the world with novel entities to account for the phantasmagorical objects of hallucination. Among the things in the representational content, some are more or less abstract than others. The typical representationalist strategy, it seems, was to consider the more abstract items as universals, and to let them have a life independently of being instantiated. That was, in very rough strokes, the strategy adopted by the conjunctivist accounts discussed in this paper.

As Thompson (2008, p. 400) warned, "we should be careful not to confuse at the outset an attractive view about the intentional content of experience with an attractive view about what metaphysically grounds the phenomenal character of an experience". The two views discussed in this paper may look quite appealing from within a certain limited set of concerns, but they may hide a much less attractive metaphysical core. According to Thompson, "we might well wonder, how can **universals** do this metaphysical job? Or, how can we be acquainted with universals in the way that seems to be required in order to account for our experience of redness?" (THOMPSON, 2008, p. 400, author's emphasis). The response typically given to these questions, alas, seems to lie far outside the naturalistic spirit that motivated representationalism in the first place. Universals presumably exist outside space-time, and they lack causal powers. The deep problem is not exactly how we come to represent universals, but rather how uninstantiated properties could ever constitute the phenomenal character. Even if universals are allowed to have a life of their own, the very relation between subjects and universals, which is constitutive of phenomenal experience, seems to be less than fully naturalistic.

The metaphysical conundrum just sketched seems to challenge Tye (2009) and Johnston (2004) alike. Direct realism, which both of them are committed to, seems to require more than they may be willing to give. The claim that properties are uninstantiated universals is a crucial one for both of them. It is hard to see how a common factor between hallucination and veridical perception could ever be found without it. But if this metaphysical problem sketched above is a genuine one, as it seems to be, conjunctivism and direct realism make strange bedfellows.

For the reasons discussed in the previous sections, strong disjunctivism seems to be equally hopeless. Strong disjunctivists have a hard time explaining the subjective indistinguishability between hallucination and veridical perception, and they don't fare any better when it comes to explaining the distinctive nature of hallucination.

We end up here left with a dilemma concerning the nature of hallucinatory experience. In this paper, I dare not fancy a solution myself. I can, however,

envision a few alternatives. One can, for instance, deny that a hallucination can possibly relate the hallucinator with novel properties. In this case, hallucinated properties could be metaphysically grounded on previously perceived instantiated properties. This line of response, however, would be committed to denying Johnston's claim that hallucinations can ground *de re* knowledge of qualities. Another alternative is to bite the bullet and affirm the universality of perceptual properties. If a detailed account of the relation between acts of awareness and universal entities can be contrived, the apparent obscurity surrounding this relation may well be dissipated.[28]

These alternatives, of course, are far from exhausting the whole terrain. As stated in the very beginning, my aim here is much less ambitious. I only aimed to compare and critically evaluate the theories of Tye (2009) and Johnston (2004), which are two influential approaches now on the market. By considering some objections to their strategy of combining weak conjunctivism and weak disjunctivism, I ended up touching some metaphysical challenges that were not, as far as I can see, appropriately addressed by any of them. I leave them here as open challenges for conjunctivists of any sort.

References

ALFORD-DUGUID, Dominic; ARSENAULT, Michael. On the explanatory power of hallucination. *Synthese*, v. 194, n. 5, p. 1765-1785, 2017.

AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and statistical manual of mental disorders:* DSM-5. 5. ed. Arlington, VA: American Psychiatric Publishing, 2013.

ARANYOSI, István. Silencing the argument from hallucination. *In:* MACPHERSON, F.; PLATCHIAS, D. (eds.). *Hallucination:* philosophy and psychology. Cambridge, MA: MIT Press, 2010. p. 255-269.

BACH, Kent. Searle against the world: how can experiences find their objects? *In:* TSOHATZIDIS, S. L. (ed.). *John Searle's philosophy of language:* force, meaning, and mind. Cambridge, UK: Cambridge University Press, 2007. p. 64-78.

BURGE, Tyler. Vision and intentional content. *In:* LEPORE, E.; VAN GULICK, R. (eds.). *John Searle and his critics*. Oxford: Blackwell, 1991. p. 195–214.

BREWER, Bill. Realism and the nature of perceptual experience. *Philosophical Issues,* v. 14, n. 1, p. 61-77, 2004.

BREWER, Bill. Perception and content. *European Journal of Philosophy,* v. 14, n. 2, p. 165-81, 2006.

BREWER, Bill. How to account for illusion. *In:* HADDOCK, A.; MACPHERSON, F. (eds.). *Disjunctivism:* perception, action, knowledge. Oxford: Oxford University Press, 2008. p. 168-180.

[28] Alford-Duguid and Arsenault (2017), for instance, take something like the former line, whereas Tye (2009; 2010) seems to take the latter.

BREWER, Bill. *Perception and its objects.* Oxford: Oxford University Press, 2011.

DANCY, Jonathan. Arguments from Illusion. *Philosophical Quarterly,* v. 45, p. 421-438, 1995.

DAVIES, Martin. Perceptual content and local supervenience. *Proceedings of the Aristotelian Society*, v. 92, p. 21–45, 1992.

DORSCH, Fabian. The unity of hallucinations. *Phenomenology and the Cognitive Sciences,* v. 9, n. 2, p. 171-191, 2010.

DUNN, Jeffrey. The obscure act of perception. *Philosophical Studies,* v. 139, n. 3, p. 367–393, 2008.

FISH, William. The direct/indirect distinction in contemporary philosophy of perception. *Essays in Philosophy,* v. 5, n. 1, 2004.

FISH, William. Disjunctivism, indistinguishability and the nature of hallucination. *In:* HADDOCK, A.; MACPHERSON, F. (eds.). *Disjunctivism:* perception, action, knowledge. Oxford: Oxford University Press, 2008. p. 144-167.

FISH, William. *Perception, hallucination, and illusion.* Oxford: Oxford University Press, 2009.

GRICE, H. P. The causal theory of perception. *Proceedings of the Aristotelian Society,* v. 35, p. 121-52, 1961.

HILBERT, David. Hallucination, sense-data and direct realism. *Philosophical Studies,* v. 120, n. 1, p. 185-191, 2004.

HUEMER, Michael. *Skepticism and the veil of perception.* Lanham, MD: Rowman & Littlefield Publishers, 2001.

HURVICH, Leo Maurice. *Color vision.* Cambridge, MA: Sinauer Associates, 1982.

LOAR, Brian. Phenomenal intentionality as the basis of mental content. *In:* HAHN, M.; RAMBERG, B. (eds.). *Reflections and replies:* essays on the philosophy of Tyler Burge. Cambridge, MA: MIT Press, 2003.

JOHNSTON, Mark. The obscure object of hallucination. *Philosophical Studies*, v. 120, n. 1, p. 113–183, 2004.

JOHNSTON, Mark. Better than Mere Knowledge? The Function of Sensory Awareness. *In:* GENDLER, T. S.; HAWTHORNE, J. (eds.). *Perceptual Experience.* Oxford: Oxford University Press, 2006. p. 260-290.

JOHNSTON, Mark. The Problem with the Content View. *In:* BROGAARD, B. (ed.). *Does perception have content?* Oxford: Oxford University Press, 2014. p. 105-137.

MARTIN, M. G. F. The limits of self-awareness. *Philosophical Studies,* v. 120, n. 1, p. 37–89, 2004.

MARTIN, M. G. F. On being alienated. *In:* GENDLER, T. S.; HAWTHORNE, J. (eds.). *Perceptual Experience.* Oxford: Oxford University Press, 2006. p. 354-410.

MCDOWELL, John. The content of perceptual experience. *Philosophical Quarterly*, v. 44, n. 175, p. 190-205, 1994.

MCGINN, Colin. *The character of mind.* Oxford: Oxford University Press, 1982.

PAUTZ, Adam. Why Explain Visual Experience in Terms of Content? *In:* NANAY, B. (ed.). *Perceiving the World.* Oxford: Oxford University Press, 2010. p. 254-309.

SEARLE, John. *Intentionality:* an essay in the philosophy of mind. Oxford: Clarendon Press, 1983.

SIEGEL, Susanna. Indiscriminability and the phenomenal. *Philosophical Studies,* v. 120, n. 1, p. 91–112, 2004.

SIEGEL, Susanna. The epistemic conception of hallucination. *In:* HADDOCK, A.; MACPHERSON, F. (eds.). *Disjunctivism:* perception, action, knowledge. Oxford: Oxford University Press, 2008. p. 205-226.

SIEGEL, Susanna. *The contents of visual experience.* Oxford: Oxford University Press, 2010.

SIMS, Andrew. *Symptoms in the mind:* an introduction to descriptive psychopathology. 2. ed. London: WB Saunders, 1995.

SLADE, P. D.; BENTALL, R. P. *Sensory deception:* a scientific analysis of hallucination. London: Croom Helm, 1990.

SMITH, A. D. Disjunctivism and Discriminability. *In:* HADDOCK, A.; MACPHERSON, F. (eds.). *Disjunctivism:* perception, action, knowledge. Oxford: Oxford University Press, 2008. p. 181-205.

SOTERIOU, Matthew. The subjective view of experience and its objective commitments. *Proceedings of the Aristotelian Society*, v. 105, p. 177–190, 2005.

THOMPSON, Brad. Representationalism and the argument from hallucination. *Pacific Philosophical Quarterly*, v. 89, n. 3, p. 384-412, 2008.

TRAVIS, Charles. The silence of the senses. *Mind*, v. 113, n. 449, p. 57-94, 2004.

TYE, Michael. *Consciousness revisited:* materialism without phenomenal concepts. Cambridge, MA: MIT Press, 2008.

TYE, Michael. The admissible contents of visual experience. *The Philosophical Quarterly,* v. 59, n. 236, p. 541-562, 2009.

TYE, Michael. What is the content of a hallucinatory experience? *In:* BROGAARD, B. (ed.). *Does perception have content?* Oxford: Oxford University Press, 2014. p. 291-310.

WILLIAMSON, Timothy. *Identity and discrimination.* Oxford: Basil Blackwell, 1990.

Recebido em: 04-03-2019

Aceito para publicação em: 25-06-19

REFLEXIVE RULES AS CONTENT: THE CASE OF DEICTIC DEMONSTRATIVES

REGRAS REFLEXIVAS COMO CONTEÚDO: O CASO DOS DEMONSTRATIVOS DÊITICOS

EDUARDA CALADO BARBOSA[1]
IIF-SADAF/CONICET – Argentina
eduardacaladobarbosa@gmail.com

ABSTRACT: Determining what content is expressed by a demonstrative when its reference cannot be determined is a problem for those who assume that demonstrative reference is cognized by interpreters and demonstrative meaning has a mere indicative role. Here, I explore a concept of content that gives meaning a cognitively relevant role, namely, John Perry's classificatory concept of content. With that purpose, I compare the interpretation of a deictic demonstrative in two cases: for an eavesdropper and a conversational participant, aiming to show that meaning, in the form of reflexive rules, can be recruited to play the role of content when information (in the speech situation) is scarce.

KEYWORDS: Demonstratives. Content. Conversation. Information. Reflexive rules.

RESUMO: *Determinar que conteúdo é expresso por um demonstrativo quando sua referência não pode ser determinada é um problema para aqueles que supõem que seu significado tem um papel meramente indicativo. Aqui, eu exploro um conceito de conteúdo que dá ao significado um papel cognitivamente relevante, a saber, o conceito de conteúdo classificatório de John Perry. Com tal propósito, eu comparo a interpretação de um demonstrativo dêitico em dois casos: para um interceptador (eavesdropper) e um participante conversacional, objetivando mostrar que o significado, na forma de regras reflexivas, pode ser recrutado para desempenhar o papel de conteúdo sempre que a informação (na situação de fala) for escassa.*

PALAVRAS-CHAVE: Demonstrativos. Conteúdo. Conversação. Informação. Regras Reflexivas.

Paradigmatic utterance situations are characterized by the fact that interpreters in them can successfully determine the reference of singular terms. In such cases, speakers and their interlocutors collaborate in joint communicative efforts to share information about individuals. For example, if two agents who are speakers of the same natural language, A and B, wish to catch a fly in a joint effort, it is very likely that they will occasionally use language to share information about that specific fly. Suppose that A wants to refer to the fly and chooses to use the definite description 'the fly' with that purpose. In this case, A will need to provide information that allows B to determine the reference of 'the fly' in that context of use. If, for some reason, either A or B fails to execute A's plan of referring properly,

[1] Pesquisadora visitante no IIF (Instituto de Investigaciones Filosoficas) - SADAF/CONICET.

and the speaker's reference cannot be determined, we have an instance of what I will call a "nonparadigmatic communicative interaction"[2].

I will be interested here specifically in nonparadigmatic communication involving deictic demonstratives, i.e., demonstratives accompanied by ostensive acts of demonstration. Moreover, I will focus on the content expressed by such terms when interpreters fail to determine reference. My aim is to look for a concept of content that explains cognitive significance, namely, how meaning and reference are cognized by competent speakers/interpreters of a language[3]. With that purpose, I will compare eavesdroppers to conversational participants.

I will start with a general characterization of demonstratives and then proceed to discuss nonparadigmatic communication. Lastly, I will present Perry's Reflexive-Referential Theory (RRT), claiming that its concept of content accounts thoroughly for eavesdropping. According to Perry, utterances of sentences with singular terms express several truth-conditional contents, depending on the information provided by each corresponding speech situation. For instance, if the information content of a situation is enough to allow for the determination of the reference of the singular term, the content expressed will be a complete proposition or, as Perry names it, the referential content of the utterance. If, otherwise, the information content fails to provide such a degree of semantic specificity, the content expressed will be bound to the utterance and the data provided by its production. Perry calls this last kind of content *reflexive*, since it is about the utterance itself and the linguistic information it carries. In the last section of this paper, after my brief presentation of Perry's multi-content approach, I will proceed to show how RRT explains the kind of content apprehended by eavesdroppers and its cognitive significance.

1 Demonstrative content

It is widely accepted by philosophers of language and of linguistics that deictic demonstratives, like 'that' and 'this' in sentences (1), (2) and (3) below, are context-dependent expressions that serve the purpose of referring to perceptual objects[4].

(1) That is my girlfriend.

(2) This is my father.

(3) That table is nice.

[2] There is more than one way in which communication can be nonparadigmatic, but I will discuss only situations in which communication is nonparadigmatic due to failure in determining reference.
[3] I take meaning here to be the linguistic rule that guides the use of an expression, following Strawson (1950). The concept of content I will look for is token-reflexive (PERRY, 2001) and respects what Wettstein (1985) and Taylor (1995) call the *Cognitive Constraint on Semantics*.
[4] To be more precise, the referents of demonstratives do not need to be concrete objects. For example, we can talk about habits, institutions and feelings with demonstratives and, also, project this mode of designation to objects that are not in the situation, such as imaginary things.

They are used to perform acts of demonstration and are associated to linguistic behavior that involves manipulation of attention with the purpose of making individuals perceptually salient to audiences. Karl Buhler (1934), for instance, who prolixly discusses *deixis* in his theory of language, affirms that members of the class of demonstrative expressions operate as orientation signs and indicators of reference.

Another way to define demonstratives is in terms of the special property of *exaphora* (Diessel, 1999, 2006); their lexical meanings stipulate that their content is to be defined relatively to facts about the utterance, which orient the interpreter outside the discourse towards the surrounding situation. Moreover, demonstratives have syntactic functions, working, for example, as pronouns – as in (1) and (2) – or noun modifiers – as in (3).

Additionally, demonstratives execute pragmatic functions[5], being used to focus the hearer's attention on objects and locations, and on the informational flow of the ongoing discourse, as in (4).

(4) That is exactly what I mean.

Now, the semantic function of deictic demonstratives, according to Diessel, is finally to indicate distance relatively to the deictic center (typically the speaker), although their meanings carry information about qualities that orient the identification of the referent, such as proximity, animacy and humanness. All languages[6] studied by Diessel had at least two demonstratives marking points on a distance scale. Other less frequent deictic features included visibility, height etc. Qualitative features, as for example, gender and number are, notwithstanding, almost universal.

In discussing pragmatic features of demonstratives, Diessel highlights the important distinction between endophoric and exaphoric uses. Exaphoric uses require manipulation of attentional behavior towards the environment, depending greatly on extra-linguistic resources. Here, I will use the terminology *deictic demonstrative* to refer to what Diessel calls "exaphoric uses of demonstratives" and I shall focus solely on the deictic demonstrative 'that', used to refer to visual objects[7].

Deictic demonstratives are not only context-dependent, but also context-variant, given that their semantic sensitivity to situations of use determine that their contents consistently vary from one occasion to another. Recanati (2003) succinctly summarizes the properties of context-dependence (or sensitivity) and context-variance of deictic demonstratives in the fragment below:

[5] The semantic profile that philosophers and linguists often ascribe to demonstratives have to do with their pragmatic function.
[6] Diessel (1999) analyzed eighty-five different languages.
[7] For visual perception, speakers manipulate visual attentional behavior towards locations occupied by the referents. Semantic theories diverge as to whether or not identifying a visual location is the same as identifying the referent. See Kasher (1998) and Grundy (2013).

> A (disambiguated) expression is *context-sensitive* or *context-dependent* if and only if its semantic content depends upon, and varies with, contextual factors such as the speaker's intention […] The *semantic content* of an expression is that property of it which (i) must be grasped by whoever fully understands the expression, and (ii) determines the expression's extension. It can be presented as a (possibly partial, and possibly constant) function from circumstances of evaluation to extensions. The *extension* of a prima facie singular term (name, pronoun, definite description, etc.) is an individual object — the reference of the term (RECANATI, 2003, p. 14).

My proposal is to explore the concept of *demonstrative content* and discuss how it is related to linguistic rules. With that in mind, I explore semantic and pragmatic features of demonstratives, following Diessel's characterization.

Let us start by taking the example of (5) below.

(5) Daddy, that is Julie!

Suppose that Jane, a dentist assistant, overhears this utterance, which is part of a conversation between a man and his daughter in the waiting room. Jane is inside the equipment room and has no visual information regarding the speaker of (5), her interlocutor or the putative referent of 'that'. Call this situation S1.

In S1, there is no coordination of attentional behavior for many discernable reasons: firstly, because the interpreter is not in the same perceptual (visual) environment as the speaker; but also, and more importantly, because she is not a conversational participant. Jane is what Clark and Shaefer (1992)[8] call an *eavesdropper*: a listener (typically a bystander) who overhears the conversation but is not acknowledged by the participants. In S1, not only there is no *common* perceptual environment; there is no set of manifestly common assumptions. Father and daughter have an *indifferent attitude towards Jane* as well as towards what she can grasp from the utterance. Because Jane is not a part of the collaborative activity in course, her informational status is not taken into consideration in designing the conversational interaction.

The problem with conversations like S1 is that eavesdroppers apprehend linguistic information, since they are competent speakers/interpreters of a natural language, but they are not in position to take part in the specific plans of communication of which the utterances that they interpret are part. In the case of Jane, she has the necessary skills to interpret the utterance of (5) truth-conditionally, but she does not participate in the plan of referring to Julie.

In the remaining of this section, I will develop the claim that Jane, though external to the conversational setting, accesses truth-conditional content. Additionally, I compare semantic interpretation in paradigmatic situations in which the determination of reference is collaborative with the case of eavesdropping.

[8] Whose categorization is inspired by Goffman (1981).

1.1 Context as conversational common ground

The term 'context' is often understood in a vague sense, as the setting in which communicative acts take place; in this sense, it has no additional technical connotation. Context as a representation of concrete speech situations for theoretical purposes was famously systematized by David Kaplan (1989). Kaplan thought that contexts were necessary for his project of developing a logic of demonstratives, but in the end, his theory was mostly focused on the special semantic nature of other context-dependent expressions, namely, pure indexicals ('I', 'here', 'now'). Contexts are then considered necessary, for Kaplan, to explain what he calls indexical nature of context-dependent expressions.

Kaplan's choice of introducing contexts is, firstly, based on the idea that the semantics of indexicals should be accounted for by a bi-dimensional semantics in which contexts determine content (or the proposition expressed/what is said)[9], while truth-values are determined by the circumstances in which contents are evaluated (actual and contrafactual). Secondly, it is motivated by his acknowledgment that contents can be represented as intensions. So, content can be shifted by intensional operators for time and possible world, as, for example, 'in ten years', 'possibly' and 'necessarily'. What I will identify here as the Kaplanian context is also defined as narrow context, and it corresponds to the set of semantically relevant parameters for the determination of content, which include the agent, the time-space location and a possible world.

Narrow context is defined in contrast with broad (or wide) context, which, on its own turn, includes all information that is communicationally relevant to the successful completion of a speech act, including what the speaker means[10]. It is then standardly assumed that broad context differs from narrow context in that it is recruited only once the proposition expressed is determined. While narrow context is used semantically, broad context is used either post-semantically – once the proposition is determined – or pre-semantically – before the proposition is determined, to fix linguistic features.

According to Recanati (2003), philosophical literature about contextual dependence is populated with different approaches to the problem of how to represent broad context. Would its role really be limited to the pre-propositional as well as to the post-propositional stage of interpretation? If the answer is 'yes', we are left with the problem of explaining the interpretation of statements with demonstratives and some indexicals, as Recanati remarks:

> We pretend to be able to manage the situation with a narrow notion of context, the kind we choose to deal with indexicals, when, in fact, we can only determine the referent intended by the speaker (i.e. the narrow context relevant to the Interpretation of

[9] For demonstratives, the Kaplanian semantics includes a) a character, the demonstrative's conventional meaning; b) the content (the character-in-context); and c) a directing intention manifested by a demonstration. The demonstration externalizes an inner intention to refer to a perceptual object on which the speaker has focused his attention.
[10] See Grice (1969, 1986).

> enunciation) when we turn to pragmatic interpretation, relying on the broad context. (RECANATI, 2003, p.5)

Here, Recanati reinforces the idea that the requirement of automatism in the attribution of occasional value to sensitive expressions, that is, of "purity" of the context/meaning relation is satisfied only by a very limited number of expressions. As he points out, even the egocentric categories 'here' and 'now' – also 'today' – are amenable to interpretation in terms of broad context[11]. To summarize, the dispute here is between those in favor of expanding the phenomenon of context-sensitivity – moderately or radically – and those more firmly committed to minimal semantics[12].

Among the contemporary philosophers who favor the alternative of representing context in terms of broad context, abdicating this exclusive concern with sensitive expressions, is Stalnaker (1999, 2002), with his conception of *discursive context*. In his view, context is not simply a theoretical construct that explains the strict phenomenon of indexicality; it is rather the implicit or presupposed "body of information that is deemed, at a certain point, as common to the participants in the discourse" (STALNAKER, 1999, p. 98). Performance of new assertions, for Stalnaker, correspond to requests to update what is being taken for granted at a moment t. Each new assertion must satisfy certain conditions: a) its content cannot be contrary or contradictory with respect to the propositions taken as true at that moment; and b) such propositions should be taken for granted by all participants. Context, for Stalnaker, is then the dynamic *presupposed common ground* of a conversational exchange.

Now, let us return to (5) and the case of eavesdropping. It seems evident that Jane, the father and the daughter do not share a presupposed common ground of assumptions. The propositional commitments that father and daughter undertake are not manifest to Jane, and her informational status is completely ignored by them. As a result, the daughter, call her Lily, will not worry about adjusting her plan of referring to Julie to what Jane knows before the utterance of (5), nor will she be concerned with what Jane will come to know afterwards.

If Jane were acknowledged by Lily, the girl's plan of referring would probably not involve a demonstration. Compare S1 to S2. Imagine that Lily knows Jane now, and that they are talking over the phone. Lily wants to tell Jane that she has a new friend, whose name is Julie. In this situation, she might utter something like (6):

[11]For a more detailed idea of the distinction between narrow and broad context, see Corazza (2014).

[12] Following Corazza (2014) and his characterization of Minimalism, I understand by minimal semantics the thesis that context is necessary for the determination of the proposition expressed by an utterance if selected by morphemes of the sentence. According to minimalist semantics, the role of semantics is to determine the truth-conditions of well-formed sentences of a natural language in accordance with the principle of compositionality. Furthermore, in this view, pragmatics is kept out of truth-conditional interpretation. Unlike semantic interpretation, pragmatic interpretation is inferential, intentional and non-conventional.

(6) Jane, do you remember Nick's sister? She is my new schoolmate. Her name is Julie.

Here, Lily takes Jane's informational status into consideration and chooses an appropriate cognitive path that will lead her interlocutor to the referent[13]. Though fallible, her plan explores a common ground of assumptions that includes the information that Julie has a brother named Nick, who Jane is familiar with. In S2, Lily's act of referring is, in contrast to S1, sucessful. Success, in this case, is measured by the level of collaboration (in what regards information sharing) between conversational participants. I will say a few words about that in the next sub-section.

1.1.1 Conversation and collaboration

As Clark and Wilkes-Gibbs (1986) point out, in conversations, reference is established in collaboration, at least paradigmatically. The aboutness of an utterance typically becomes part of the common ground by means of coordinated behavior (acceptance, rejection, etc.) and by the systems of repair and reinforcement that permit communicational improvement whenever it is necessary:

> People in conversation manage who is to talk at which times through an intricate system of turn taking […]. Further, when one person speaks, the others not only listen but let the speaker know they are understanding-with head nods, yes's, uh huh's, and other so-called back channel responses (Duncan, 1973; Goodwin, 1981). When listeners don't understand, or when other troubles arise, they can interrupt for correction or clarification […]. The participants also have techniques for initiating, guiding, and terminating conversations and the topics within them […] (CLARK & WILKES-GIBBS, 1986, p. 2).

Luckily for speakers, linguistic communication is the fluid, self-preserving process described above. We have mechanisms to accommodate or exclude non-collaborative conversational contributions and recuperate the rational character of paradigmatic communication. Their idea is that all participants are mutually responsible for establishing what was said, mostly because they rely on shared information all the time, and they make assumptions about each other's statuses. Also, speakers need the repair mechanisms because accomplishing plans of referring depends greatly on how well the interlocutor is following every step, grasping every piece of information delivered by the speaker.

In our example of S1, Jane is not acknowledged by the other participants. So, there can be no division of labor in establishing what was said. Additionally, Jane, an eavesdropper, cannot take part in repair mechanisms in a pragmatically

[13] The contrast here – in the cases of (5) and (6) – is between a demonstrative and a rich description of the referent. The idea is to compare two modes of designation: demonstrating and describing.

appropriate way. Eavesdroppers are excluded from what Clark & Wilkes-Gibbs call acceptance cycles.

> The basic process, which might be called the acceptance cycle, consists of a presentation plus its verdict. Let x, y, and z stand for noun phrases or their emendations. A presents x and then B evaluates it. If the verdict is not positive, then A or B must refashion that presentation. That person can offer: a repair x', an expansion y, or a replacement z. The refashioned presentation, whether x', x + y, or z, is evaluated, and so on. Acceptance cycles apply iteratively, with one repair, expansion, or replacement after another, until a noun phrase is mutually accepted. With that, A and B take the process to be complete (CLARK & WILKES-GIBBS, 1986, p. 24).

A request of repair from Jane's part, in trying to determine the referent of 'that' in (5), would doubtlessly be taken as an inadequate behavior. All information Jane will be able to gather from the utterance will have to respect her status of non-participant in the conversation. So, in S1, she will fail to apprehend the content that Lily intended to express to her father. But, what content does she apprehend, as a competent speaker of English?

Lily and her father in S1 are exchanging information about the individual Julie. If Lily's plan of referring succeeds, he will be able to identify the *demonstratum* of 'that', with the help of some ostensive act that makes Julie salient. The conventional meaning of 'that' will accomplish its task in indicating which individual in the perceptual environment it denotes, if and only if it offers the adequate conditions of identification in that context of use. In a paradigmatic situation, Lily's father will apprehend the following truth-conditions, where the boldface marks the contribution made by the demonstrative to the proposition expressed:

(P5) **Julie** is named 'Julie'.

However, (P5) is not what Jane apprehends. Because she has scarce information in S1, all she has at her disposal are the conventional meanings of the expressions used by Lily in uttering (5), and the little she knows about the production of that token. Thus, it seems that Jane accesses something like (P^r5), where the italics mark the conditions of identification carried by the conventional meaning of 'that':

(P^r5) That *the individual made salient by the speaker of (5), standing at distance* d *relatively to the speaker,* is named 'Julie'.

Now, compare S1 with a new situation, S3. In S3, Jane sees the speaker, the interlocutor and the other girl, who is the referent of 'that'. In this communicative situation, Jane accesses (P5). She identifies the individual which the utterance of (5) is about and the conditions in which that utterance is true. In contrasting S3

and S1, we see that in the new situation, Jane, though still a bystander, whose informational status is not a part of the common ground, can successfully identify the aboutness of the token. More, though absent from the speaker's communicative plan, Jane is now in the same perceptual environment as the speaker, and she can adjust her attention according to the speaker's indications, even if she is not asked to do so.

(P^r5) is called by John Perry (2001) the *reflexive truth-conditions* (of (5)). Perry defines it as not being about the referent of the singular term in the sentence, but about the utterance itself. According to his theory, content is a resource used by rational agents to classify states of affairs in terms of conditions of success, so truth-conditional content is what explains, among other things, the adequacy of beliefs to evidence. I will now explain Perry's notion of content and return once again to the analysis of (5), before I advance a few observations on demonstrative content.

2 Perry's concept of content

Consider the example, in Perry (2001), of a man who sees the daily newspaper on his balcony one morning and acquires the belief that the newspaper is on the balcony. The content of the belief classifies the state of affairs, given some facts about the world and the information that is perceptually accessible to the agent. A neighbor who observes the situation and sees the man pick the newspaper from the floor is justified in assigning to him the belief that the newspaper is on the floor and, furthermore, to suppose that his action can be explained by the possession of such a belief.

In Perry, content is also goal-oriented. Take a variant of his original example. Suppose that the man picking the newspaper from his balcony floor that morning is named John, and he is an old acquaintance of Louie, the newspaper delivery man. John just found out that his and Louie's old friend, Joe, passed away, and he wishes to tell Louie the sad news. John has expectations about such conversation, but when he goes out to the balcony with the purpose of waiting for Louie, he sees that the newspaper was already delivered. What beliefs can we attribute to John? We can reliably say that John believes that the newspaper is on the floor. Nonetheless, given our narrative and John's goals, expectations and what is perceived by him, it is also reasonable to suppose that John has the belief that Louie has already delivered his newspaper that morning. Notice that the content of John's belief depends on what John's classificatory targets are. That is why Perry argues that the classificatory concept of content strongly depends on the *information content available* in the situation that the agent classifies.

Perry holds that his concept of content is particularly advantageous for an adequate theory of language because it accounts for how human language is used to communicate cognitive states, given that such states ultimately influence the way we act. As a matter of fact, the central claim in Perry (2001) is that there is a system

of truth-conditional contents at the theorist's disposal to explain how linguistic and information content are adjusted to one another. He points out that:

> […] the concept of 'truth-conditions of an utterance' is a *relative concept*, although it is often treated as if it were absolute. Instead of thinking in terms of the truth-conditions of an utterance, we should think of *the* truth-conditions of an utterance *given* various facts about it. And when we do this we are led to see that talking about *the* content of an utterance is an oversimplification (PERRY, 2001, p. 80).

Consequently, he heavily criticizes what he calls *the principle of the unique content*. Perry claims that more than one semantic content is generated by utterances of declarative sentences (with referential expressions). One of them is the *proposition expressed*[14]: one that involves a n-ary relation between a property and n individual(s), <*Ix*; *e* >, which must, in some way, relate to the conditions that make the utterance true. But, still, why should the proposition be considered a kind of content or, better still, why has it been equated with the concept of content by an important tradition in the philosophy of language?[15].

According to Perry, the main reason why the proposition is equated with the notion of content is its situational specificity and the fact that it is what typical agents *say* in typical communicative interactions. Propositions represent a high level of semantic specificity, because an interpreter only apprehends it once she loads all the spatial-temporal and intentional aspects that are relevant to semantic interpretation into truth-conditions, leaving no room for ambiguity or unfilled semantic slots. It is supposed to be the result of semantic interpretation, the content that the speaker wished to convey by performing her speech act in the first place. But, as he insists, the proposition is not the only content that is conveyed. It is given, partially, by the conventionally meaningful parts of the sentence, but more fundamentally, by the relation that the language associates with sentences, contextual factors and circumstances. He affirms: "The truth-conditions of an utterance derive directly from the meaning assigned to the sentence involved, whereas which proposition is expressed depends also on the agent, time and circumstances of utterance". (PERRY, 1997, p. 197). Perry, however, takes nonparadigmatic communicative situations as *explananda*. It is because his approach on content is, in a sense, instrumentalist at heart, allowing for multiple possibilities of interpretation, that he can account for such situations.

According to the Perryan framework, the system of contents expressed by the utterance of (5), then, will include (P^r5) and (P5). In S1, however, Jane, being an eavesdropper, will apprehend (P^r5), truth-conditions that are about the conventionally meaningful parts of (5) – as well as about the production of the utterance itself –, and not (P5). That will be the case because the content expressed by (5) in S1 is relative to the information content of the situation. But one important

[14] Though he is not exactly concerned with the ontology of propositions, he takes the classic characterization of propositions as *sharable objects of belief*, the semantic content to which we attribute truth-values.
[15] Of which Kaplan is an example.

thing is still left to be explained: what is the role of reflexive truth-conditions like (P^r5) in explaining interpretation? Can we affirm that conventional meaning has a cognitively significant role in cases such our example or does demonstrative meaning "drops out of the picture" (NUNBERG, 1993) to give room to content?

To answer this question, let us recap some of our previous conclusions about the case of S1. Firstly, S1 was characterized as the setting of a nonparadigmatic communicative interaction in which the interpreter of the utterance of (5) is an eavesdropper, that is, an overhearer who is not acknowledge by conversational participants. I defined participants according to a common ground notion of context, based on Stalnaker's seminal ideas about the representation of context. Next, I completed this definition with some observations about how participants contribute to conversations, using Clark & Wilkes-Gibbs idea of cycles of acceptance. Then, I set to look for a concept of content that could explain what eavesdroppers apprehend as interpreters of a natural language.

Lastly, I outlined Perry's concept of classificatory content and his critique to the principle of unique content. According to him, one utterance expresses virtually several contents, depending on the information carried by the speech situation. In this view, then, an utterance of (5) in S1 expresses a content that is about linguistic information: about the utterance and its production. We seem thus to have found an answer to one of our first questions: what content does an eavesdropper apprehend?

Following Perry, I suggested that such content corresponded to the reflexive truth-conditions of (5), namely (P^r5). In effect, what seems to be at issue in the case of S1 is to provide a distinction between the content apprehended by a conversational participant and the content apprehended by a sheer interpreter. In Jane's case specifically, because the demonstrative in (5) is a context-dependent expression that requires either that the speaker and the interpreter are in the same perceptual environment, or that they collaborate in the same plan of referring, the information content of S1 is too scarce to allow for the determination of reference. Perry's reflexive truth-conditions come in handy precisely to explain interpretation situations such as S1.

Nonetheless, reflexive content is about the token of a sentence and the linguistic conventions it incorporates. In the case of a deictic demonstrative, it is about the conditions of identification that work as individuating properties relatively to contexts. These conditions have the form of a linguistic rule, but they also depend on the speaker's ability to make the putative referent salient to her audience. Both aspects have an indicative function, in the sense that they serve as pointers to individuals, given their spatial locations. The linguistic rule, however, underspecifies truth-conditions, because it does not provide a property that uniquely identifies the referent, as content is, *prima facie*, supposed to. After all, the reference of 'that' will always depend greatly on intentional elements, and the demonstrative's meaning seems to "drop out of the picture" once content is determined.

My second question concerns this view on the dichotomy between demonstrative meaning and demonstrative content in what concerns cognitve significance. It seems certainly correct that meaning alone, in the case of a context-dependent and context-variant expression, such as a deictic demonstrative, underspecifies reference. Recanati (2004) presents some convincing arguments to this effect. Yet, not all cognitively significant tasks in the interpretation of a deictic demonstrative have to be executed by the final content, that is, the proposition. In determining *how* one says something in using the word 'that', conventional meaning leaves an informational trace that can be used by competent interpreters in reconstructing what was conveyed by the speaker in a given situation.

Remember that, in S1, the reflexive linguistic rules involved in the utterance of (5) were the only sources of information at Jane's disposal in adjusting her belief-states to what she overhears. In the event of being asked about what was said by (5), Jane would probably recruit her classificatory practices to generate content like (P^r5). Our example seems to show, then, that, when communication is somehow defective, information about the utterance may come to the rescue of informativeness in communication, allowing for the rational reconstruction of (linguistic) behavior.

Furthermore, classificatory practices play exactly this role: permitting that agents harness information following requirements of success and rationality, even in situations with informational scarcity in what concerns the determination of reference. The classificatory concept of content helps us explain these very mundane situations of information harnessing. In accepting this claim, we should keep in mind that, whenever the informational game of language is ineffectual, defective, abnormal, whatever allows speakers to play the game, will do the trick of taking on the role of content.

References

BüHLER, Karl. *Sprachtheorie*. [*S.l.:s.n.*], 1934.

CORAZZA, Eros. Context, non-specificity and minimalism. *Manuscrito.* Campinas, v. 37, n. 1, pp. 5-47, 2014.

CLARK, Herbert. H. *Arenas of language use*. Chicago: University of Chicago Press, 1992.

CLARK, Herbert. H., and WILKES-GIBBS. Deanna. Referring as a collaborative process. *Cognition*. Amsterdã, v. 22, n. 1, pp. 1-39, fev. 1986.

DIESSEL, Holgar. *Demonstratives: Form, function and grammaticalization.* Amsterdã: John Benjamins Publishing, 1999.

DIESSEL, Holgar. Demonstratives, joint attention, and the emergence of grammar. *Cognitive linguistics*. Berlim, v. 17, n. 4, pp. 463-489, dez. 2006.

GOFFMAN, Erving. *Forms of talk*. Filadélfia: University of Pennsylvania Press, 1981.

GRICE, H. Paul. *William James Lectures*. Cambridge MA: Harvard University Press, 1967.

_____. *Studies in the way of words*. Cambridge MA: Harvard University Press, 1989.

GRUNDY, Peter. *Doing pragmatics*. Londres: Routledge, 2013.

KAPLAN, David. 'Demonstratives'. In: ALMOG, Josef, PERRY, John, and WETTSTEIN, Howard (eds.). *Themes from Kaplan*. Oxford: Oxford University Press, 1989.

KASHER, Asa. *Pragmatics: Critical concepts: presupposition, implicature and indirect speech acts*. Londres: Routledge, 1998.

KORTA, Kepa, and PERRY, John. *Critical pragmatics: An inquiry into reference and communication*. Cambridge: Cambridge University Press, 2010.

NUNBERG, Geoffrey. Indexicality and deixis. *Linguistics and philosophy*. Dordrecht, v. 16, n. 1, pp. 1-43, fev. 1993.

PERRY, John. *The problem of the essential indexical: and other essays*. Oxford: Oxford University Press on Demand, 1997.

_____. *Reference and reflexivity*. Stanford: CSLI, 2001.

RECANATI, François. What is said and the semantics/pragmatics distinction. *In*: BIANCHI, Claudia. e PENCO, Carlo (eds). *The Semantics/Pragmatics distinction: Proceedings from WOC 2002*. Stanford: CSLI Publications, 2003. Disponível em: https://jeannicod.ccsd.cnrs.fr/ijn_00000374/document. Acesso em: 20 maio de 2019.

_____. Indexicality and context-shift. *Workshop on indexicals, speech acts and logophors*. Cambridge MA: Harvard University, 2004.

STALNAKER, Robert. *Context and content: Essays on intentionality in speech and thought*. Oxford, Oxford University Press, 1999.

_____. Common ground. *Linguistics and Philosophy*. Dordrecht, v. 25, n. 5-6, pp. 701-721, dez., 2002.

STRAWSON, P. F. On referring. *Mind*. Oxford, New Series, v. 59, n. 235, pp. 320–44, jul. 1950.

TAYLOR, Kenneth Allen. Meaning, reference and cognitive significance. *Mind and Language*. New Jersey, v. 10, n. 1-2, pp. 129-188, mar. 1995.

WETTSTEIN, Howard. Has semantics rested on a mistake? *The journal of philosophy*, v. 83, n. 4, pp. 185-209, abr. 1986.

Recebido em: 06-03-2019

Aceito para publicação em: 25-06-19

A CRITICAL APPROACH TO SENSORIMOTOR CONTINGENCY THEORY: BRAIN AS AGENT AND CONSCIOUS MIND AS A GUIDE OF ACTION

UMA ABORDAGEM CRÍTICA À TEORIA DA CONTINGÊNCIA SENSORIOMOTORA: O CÉREBRO COMO AGENTE E A MENTE CONSCIENTE COMO GUIA DE AÇÃO

JONAS GONÇALVES COELHO[1]
Universidade Estadual Paulista (UNESP) – Brasil
jonasgcoelho@gmail.com

ABSTRACT: I present and consider critically O'Regan and Noë's sensorimotor contingency theory, proposed as an alternative to solve the explanatory gap problem. I start with the criticism that these authors address the current conception of representation, according to which conscious experiences are representations of the external world produced by the brain. Afterward, I summarize the way the sensorimotor contingency theory addresses the problem of the explanatory gap, explaining the existence, form, and content of visual consciousness in terms of an "exploratory activity" mediated by sensorimotor contingency laws. Finally, in agreement with criticisms addressed to O'Regan and Noë's solution, I propose a way to face the problem of the explanatory gap, which, recognizing the relevance of the body and the external environment to the existence, form and content of visual consciousness, but privileging the role of the brain as an organ of visual consciousness, and as an agent who uses visual consciousness as a guide to initiate and maintain embodied and situated adaptive actions in the world.

KEYWORDS: O'Regan and Noë. Sensorimotor contingency theory. Explanatory gap. Qualia. Agent brain. Consciousness guide of action.

RESUMO: *O objetivo é apresentar e refletir criticamente sobre a teoria da contingência sensoriomotora proposta por O'Regan and Noë para resolver o problema da lacuna explicativa. Começo pela crítica que esses autores dirigem à concepção representacionista corrente segundo a qual as experiências conscientes seriam representações do mundo externo produzidas pelo cérebro. A seguir, apresento, resumidamente, o modo como a teoria da contingência sensóriomotora enfrenta o problema da lacuna explicativa explicando a existência, forma e conteúdo da consciência visual em termos de uma "atividade exploratória" mediada pelas de leis da contingência sensoriomotora. Por fim, em acordo com as críticas de alguns comentadores à solução proposta por O'Regan and Noë, aponto um caminho para enfrentar esse problema, o qual, embora ressaltando o papel indispensável do corpo e do ambiente externo ao corpo na geração das formas e conteúdos da consciência visual, privilegia o cérebro como órgão da consciência visual e da mente consciente em geral, e como o agente que usa sua consciência visual como guia para iniciar e manter ações adaptativas no ambiente em que vive.*

PALAVRAS-CHAVE: *O'Regan and Noë. Teoria da contingência sensório-motora. Lacuna explicativa. Qualia. Cérebro agente. Consciência guia de ação.*

[1] Departamento de Ciências Humanas da UNESP de Bauru e Programa de Pós-Graduação em Filosofia da UNESP de Marília.

1 Traditional approach to visual consciousness, qualia and the explanatory gap

1.1

In a challenging paper published in 2001, followed by peer commentaries and authors' responses, "The sensorimotor account of vision and visual consciousness", Kevin O'Regan and Alva Noë propose to answer the following question: "What is visual experience and where does it occur?" (2001, p. 939). From the beginning and throughout the article, they criticize what would be the current neurophysiological, psychophysical, and psychological approach to vision, that is, the idea that "when we see, the brain produces an internal representation of the world" and it is the "activation of this internal representation" that is supposed "to give rise to the experience of seeing" (2001, p. 939). More precisely, O'Regan and Noë criticize both ideas, that "somewhere in the brain an internal representation of the outside world must be set up which, when it is activated, gives us the experience that we all share of the rich, three-dimensional, colorful world" (2001, p. 939), and that "cortical maps - those cortical areas where information seems to be retinotopically organized – might appear to be good candidates for the locus of perception." (2001, p. 939). O'Regan and Noë also refer to this view saying that it is a "theory of vision in which there is a picture-like internal representation of the outside world" (2001, p. 953), or "an internal, more or less picture-like, representation of the visual world." (2001, p. 955).

O'Regan and Noë consider this view as part of a broader approach to the relationship between brain and consciousness, according to which "consciousness is an *intrinsic* property of neural states", which would have "an additional property of being *phenomenologically* conscious." (2001, p. 965). In other words, a "set of neurons" or "neural representations" would correlate "strongly with aware perceptual states", and this would happen "because these neurons are probably linked to the mechanisms that are generating awareness" (2001, p. 966). Thus, "the discovery of *perfect* correlation would give us reason to believe that we had discovered *the* neural activity sufficient to produce the experience." (2001, p. 967). From this view of the brain-consciousness relationship, the problem of consciousness would be "to understand what processes or mechanisms or events in the brain make certain contents phenomenologically conscious" and "where, and how, does consciousness happen in the brain." (2001, p. 965).

O'Regan and Noë do not think that this is the main problem of consciousness, at least, from a philosophical perspective. Taking visual consciousness as a paradigmatic example, they agree that there are cortical maps, retinotopically organized, which contain information about the visual world. However, its presence and particular organization "can neither *in itself* explain the metric quality of visual phenomenology" nor "why an activation of cortical maps should produce visual experience." (2001, p. 939). O'Regan and Noë believe that a perfect correlation between visual consciousness and a set of neurons would just show that the "neural activity played some role in vision", whithout explaning why or how this neural activity produces that particular experience:

> suppose we were to discover that in the pineal gland of macaque monkeys there was a tiny projection room in which what is seen by the monkey was projected onto an internal screen whose activity correlated perfectly with the monkey's visual awareness. On reflection it is clear that such a discovery (which would surely be the Holy Grail of a neural correlate of consciousness seeker!) would not bring us any closer to understanding how monkeys see. For we would still lack an explanation of how the image in the pineal gland *generates* seeing; that is, how it enables or controls or modulates the forms of activity in which seeing consists. (2001, p. 966-967).

O'Regan and Noë refer to the problem of brain-consciousness relationship as the "qualia problem", being qualia "frequently characterized as the 'phenomenal', or 'qualitative,' or 'intrinsic' properties of experience", known through introspection and being "typically contrasted with 'intentional' or 'representational' or 'functional' features." (2001, p. 960). Considering roughly that qualia would be the ways in which we experience something subjectively, the qualia problem, also known as the "explanatory gap", would consist in connecting subjective and objective states/processes: "It has been suggested on this point that there is an unbridgeable 'explanatory gap', that it is not possible to explain the subjective, felt aspects of experience in behavioral, physical or functional terms" (2001, p. 960). In a more recent paper, "Sensorimotor theory of consciousness", published in 2015, Kevin O'Regan and Jan Degenaar, quoting Joseph Levine, refer to the explanatory gap in somewhat different terms. Adding the notion of "description", they say that the "challenge in explaining the quality of experience is to avoid an 'explanatory gap' between descriptions of the biological or physical processes involved in experience and descriptions of the phenomenal quality of experience." (2015, p. 2).

1.2

To avoid the qualia problem, or the explanatory gap, O'Regan and Noë argue that qualia, understood as "properties of experiential states or events" (2001, p. 960) or occurrences, do not exist. However, by denying the existence of qualia as just defined, the authors "are not denying that experience has a qualitative character" (2001, p. 960), but a sort of passive way of thinking about it: "Our claim, rather, is that it is confused to think of the qualitative character of experience in terms of the *occurrence* of something (whether in the mind or brain)." (2001, p. 960). O'Regan and Noë say both, that "scientists and philosophers frequently get the phenomenology of experience wrong; they misdescribe what perceptual experience is like" (2001, p. 960), and that "by denying the need for *qualia* we are not denying the existence of perceptual experience, or the possibility of phenomenological reflection on experience." (2001, p. 971).

Consistent with this claim O'Regan and Noë declare their great sympathy for the phenomenological tradition of Husserl and Merleau-Ponty, who presented "a clear and rigorous conception of the methodology of first-person investigations of experience" (2001, p. 973), and "make contributions toward the development of

a first-person study of consciousness which does not rely on the problematic conception of qualia criticized above" (2001, p. 973). Thus, O'Regan and Noë argue for "a more full-blooded phenomenological project" (2001, p. 973), defending a phenomenological approach which they believe "provides an account of the subject matter of phenomenology that is superior to that put forward by qualia-oriented positions" (2001, p. 962). Superior in two fundamental aspects: first, because it would be a theory "supported by careful reflection on what it is like to have perceptual experience" (2001, p. 962), which would exclude any conception that assumes the existence of a "detailed internal representation of the environment in the head." (2001, p. 962); second, differently from "traditional qualia-based approaches to experience", which "threaten to make experience itself something mysterious and inaccessible", the account to be proposed "helps place phenomenology as an undertaking on solid ground." (2001, p. 962).

But what rigorous description of experience would result from a phenomenology without qualia, as defined above? Before presenting O'Regan and Noe's answer to this question, it is necessary to clarify their use of the terms "consciousness" and "awareness", in order to understand what they mean by subjective and qualitative aspects of experience. By defining consciousness as *transitive consciousness*, or *consciousness of*, from the visual consciousness case, the authors say that being transitively conscious is to be *aware* of a feature of a scene, meaning that it is not an unconscious automatic exercise, rather an activity involving attention, in the sense that it is integrated in the "current planning, reasoning and speech behavior". (2001, p. 960). O'Regan and Noë's example of a driver who drives a car while talking to a friend helps to clarify this point. When driving the car, although his brain is tuned to sensorimotor contingencies related to both the relevant features of the scene and his driving behavior, such as steering and speed adjustment, by talking to his friend, the driver would not be *aware* of most of those features. In other words, the driver would be acting as an automatic pilot controlling the flight of an airplane, that is, his behavior would be regulated by appropriate sensorimotor contingencies, but he would be visually unaware of the relevant characteristics of the scene. Differently, visual *awareness* comes into the picture when *conscious* attention for thought and planning is present: "But if you should turn your attentions to the color of the car ahead of you, and think about it, or discuss it with your friend, or use the knowledge of the car's color to influence decisions you are making, then, we would say, you are aware of it." (2001, p. 944).

After arguing that O'Regan and Noë use the terms "awareness" and "consciousness" meaning the qualitative and subjective aspects of conscious experience, I return to the question of the explanatory gap from the concept of qualia. First, as it was said before, O'Regan and Noë deny the existence of qualia, that is, they criticize the description of conscious experience intrinsic to the notion of qualia, and, consequently, the existence of an explanatory gap: "there is no explanatory gap because there is nothing answering to the theorist's notion of qualia. That is, we reject the conception of experience that is presupposed by the problem of the explanatory gap." (2001, p. 962). Second, as it was also said before,

despite O'Regan and Noë´s denying the existence of qualia, they argue that conscious experience has a qualitative character: "we can defend this claim even though we do not deny, as we have been at pains to explain above, that there are experiences and that experience has qualitative character." (2001, p. 963). Thus, the authors propose a "phenomenological reflection on experience" which "captures what we believe, as experiencers, about our experiential life, but does so in a manner that does not give rise to the mystery of the explanatory gap." (2001, p. 971).

Assuming, without further justification, that the traditional representationist approach creates, but does not solve, the explanatory gap problem, O'Regan and Noë claim to solve this problem with what they call "sensorimotor theory of consciousness", also termed "sensorimotor contingency theory." As it will be seen in the next section, it is with this theory that O'Regan and Noë try to explain both, the *existence of particular forms of consciousness*, taking visual consciousness as a paradigmatic instance, and the *existence of consciousness in general*, thus facing this "absolute question" from "comparative questions" such as: "What explains that some environmental properties are consciously experienced while others are not? What explains that we sometimes are conscious while in other cases we are not (e.g. knocked out)? What explains that some systems (e.g. humans) have conscious experience while others (e.g. thermostats) do not?" (2015, p. 4).

2 The sensorimotor contingency theory and the explanatory gap

2.1

To overcome the explanatory gap, O'Regan and Noë propose, against the current "mysterious assumption", a "natural way" of addressing the "problems about the nature of visual consciousness, the qualitative character of visual experience, and the difference between vision and other sensory modalities." (2001, p. 940). Their view is centered on the idea that vision is an "exploratory activity" mediated by "sensorimotor contingencies": "*a mode of exploration of the world that is mediated by knowledge of what we call sensorimotor contingencies.*" (2001, p. 940). What O'Regan and Noë mean by "sensorimotor contingency," is the "*structure of the rules* governing the sensory changes produced by various motor actions." (2001, p. 941). So, to have visual perception is the same as "being able to exercise control of the rules of sensorimotor contingency associated with vision" (2001, p. 943). These rules are also referred to as "laws".

The rules/laws related do visual consciousness would be both, those inherent to the body and those involving the relationship between the body and the outside environment. There is a rule/law governing the relationship between eye movements and retinal events: "when the eyes rotate, the sensory stimulation on the retina shifts and distorts in a very particular way, as determined by the size of the eye movement, the spherical shape of the retina, and the nature of the ocular optics." (2001, p. 941). Another rule/law relates body movement to the pattern of flow on the retina: "the flow pattern on the retina is an expanding flow when the body moves forwards and contracting when the body moves backwards." (2001,

p. 941). There is also a rule/law governing the relationship between the closing of the eyes during blinks causing the "stimulation to change drastically, becoming uniform (i.e. the retinal image goes blank)." (2001, p. 941). But the constitutive rules/laws of sensorimotor contingencies are not limited to the relations between the bodily components relevant to the visual consciousness; they also involve the properties of the objects and external environment. Thus, features such as size, shape, texture and color, as well as distances and angles relative to the observer, are objective aspects that should be considered as fundamental for visual consciousness, considering the limits of the constitutive information of the retinal image and its dependence on those elements external to the retinal image.

However, these rules/laws related to the body and the external environment would just be necessary conditions, not being sufficient to explain the conscious aspect of vision, which would also depend on the existence of purposes of "thinking", "planning", "reasoning" and "action guide". This third essential component for the existence of visual consciousness is presented by O'Regan and Noë with the example of driving a car. The idea is that when driving we are faced with a scenario whose sensorimotor contingencies are partly used to control our behavior, for example, to adjust the direction or speed of the car. It happens that, although we have our behavior regulated by this type of sensorimotor contingencies, we are often, in relation to them, as "automatic pilots controlling the flight of an airplane" (2001, p. 944), that is, we remain "visually unaware of the associated aspects of the scene." (2001, p. 944). Hence, in order to have visual awareness, it is necessary, "in addition to exercising the mastery of the relevant sensorimotor contingencies, to make use of this exercise for the purposes of thought and planning" (2001, p. 944), or, as it is also said by the authors, for the purposes of thought, control, reasoning and action. These three essential aspects of visual consciousness are summarized by O'Regan and Noë:

> one important dimension of what it is like to see is fixed by the fact that there is a lawful relation of dependence between visual stimulation and what we do, and this lawful relation is determined by the character of the visual apparatus. A second crucial feature that contributes to what it is like to see is the fact that objects, when explored visually, present themselves to us as provoking sensorimotor contingencies of certain typically visual kinds, corresponding to visual attributes such as color, shape, texture, size, hidden and visible parts. Together, these first two aspects of seeing, namely, the visual-apparatus-related sensorimotor contingencies, are what make vision *visual*, rather than, say, tactile or auditory. Once these two aspects are in place, the third aspect of seeing, namely, visual awareness, would seem to account *for just about all the rest* of what goes into making up the character of seeing. For, visual awareness is precisely the availability of the kinds of features and processes making up the first two aspects for the purposes of control, thought, and action. (2001, p. 944).

But, how this approach, involving the three mentioned aspects of visual consciousness, contributes to overcoming the explanatory gap requires further explanation. O'Regan and Noë try to do that in the authors' response to open peer

commentary, where they reply to the criticisms "by formulating more explicitly the reasoning implicit in the target text." (2001, p. 1011). There, they address the question, which they call "basic", that is, "can the sensorimotor approach explain why activity drawing on knowledge of sensorimotor contingencies gives rise to experience at all?" (2001, p. 1011). Their first step is to explain what they mean by the qualitative aspect of visual experience, or the what it is like to have a visual experience, in the sense proposed by Thomas Nagel in the article "What is it like to be a bat?". They consider that the defining features of the qualitative aspect that need to be explained by the sensorimotor approach are the following: visualness, forcible presence, ongoingness and ineffability. Let's see how O'Regan and Noë define and explain each one of these features of visual conscious experience.

Regarding the first aspect, that is, the visualness of conscious experience, the authors believe that what makes it visual and non-auditory, olfactory, tactile, etc., as I already said in the second section of this paper, is the specific sensorimotor contingencies "mediated by the visual apparatus and by the character of the sensory changes produced by objects as they move in space." (2001, p. 1012). It can be inferred from this that something similar happens with conscious auditory, olfactory, tactile, etc. experiences, with the difference that the mediation would involve other sensory and environmental devices related to these different modalities.

The second aspect of visual conscious experience, the forcible presence, is explained by two notions: grabbiness and bodiliness. Grabbiness means that an object forces its presence to become conscious to the perceiver, thus attracting his attention. Bodiliness means that the sensory stimulation coming from an object changes as we move the whole body, or parts of it, such as the eyes and head, relevant to the perception of that object. Grabiness and bodiliness would also explain the third aspect of conscious experience, that is, the feeling of an ongoing qualitative state. The awareness of experience as ongoingness would result from the fact that sensory events are always present when we look at them, that is, when they attract our attention, and vary regularly according to our bodily movements.

The fourth aspect of visual conscious experience to be explained is the fact that conscious experience appears to us as ineffable. This sense of ineffability would be the result of our ignorance on the complex sensorimotor laws that govern it, for example, the complex processes involving the functioning of the eye and its relation to visual stimuli are not consciously available to us, thus impossible to be described, although possible to be used: "our sense of the ineffability of experience is explained by the fact that we lack access to the very complicated laws governing the sensorimotor contingencies involved in sensorimotor exploration." (2001, p. 1012).

I would like to finish this section by making a brief comment about the role of the brain in the sensorimotor contingency theory. By distinguishing the visual contingencies related to the visual apparatus and objects from the processes of thought, planning, reasoning and control, O'Regan and Noë mention the involvement of the brain only with the first two necessary but not sufficient elements of visual consciousness. They say that the function of the brain is to

encode the visual attributes according to the vision-specific laws of sensorimotor contingency. In the case of the visual quality of shape, for example, the brain would abstract from the infinite "set of all potential distortions that the shape undergoes when it is moved relative to us, or when we move relative to it […] a series of laws, and it is this series of laws which codes shape." (2001, p. 942). The same reasoning is applied to other kinds of visual quality, it being the role of the brain, in this practical and non-propositional knowledge, to extract the laws that are archived and applicable whenever new visual stimuli are present.

Thus, in specifying the role of the brain in visual consciousness, O'Regan and Noë restate their position critical of the traditional neuroscientific view according to which the content of visual experience would be generated by the activation of specific neural substrates. Strictly speaking, there would be no neural substrates specific to vision, since the specification of visual contents would depend on the knowledge of sensorimotor contingencies involving the interaction between the visual apparatus and the environment. By assuming that conscious visual experience consists of an activity of exploring the environment, the authors argue that visual experience is not an "occurrence" derived from neural activity of the brain, the role of this organ being only "the mastery and exercise of the laws of the sensorimotor contingency." (2001, p. 968). Thus, the brain is considered just as "an element in a system, and not, as it were, as the seat of vision and consciousness all by itself." (2001, p. 970).

2.2

O'Regan and Noë believe that the sensorimotor contingency theory above summarized, is good enough to explain the qualitative character of conscious experience, that is, to explain why a sensation has a feel and why this feel has particular features in each distinct sensation, thus solving the problem of the explanatory gap: "that is, the problem of explaining perception, consciousness, and qualia in terms of physical and functional properties of perceptual systems" (2001, p. 1020). But, for reasons such as those we will see next, some philosophers do not agree with O'Regan and Noë, arguing that their view is not very successful in overcoming the explanatory gap problem.

As it was said at the beginning of the first section, the 2001 paper by O'Regan and Noë was published along with numerous comments, many of them objections, and the authors' responses. Here I highlight just a few among the many critical comments, namely those centered on the notion of qualia and the problem of the explanatory gap. Regarding the question of qualia, a fundamental point concerns how O'Regan and Noë understand it. As it was shown in the first section, they do not accept the existence of qualia, defined as internal conscious subjective states/events/occurrences produced by the brain, although they defend the existence of conscious subjective experiences, which have a qualitative character and are amenable to a phenomenological description. Thus, it can be stated that the sensorimotor contingency theory is not a view of the mind-body relationship that proposes to reduce conscious mind to physical and/or functional and/or

behavioral events/processes. And it is from this interpretation that some critics argue that O'Regan and Noë have failed in their effort to overcome the explanatory gap, as will be seen next.

Andy Clark and Josefa Toribio, for example, agree with O'Regan and Noë's emphasis on the intimate relationship between conscious content and embedded action, but disagree that they have succeeded in dissolving or avoiding the "hard problem". Thinking of the case of vision, it could be said that a robot that is an excellent ping-pong player would meet the conditions established by the sensorimotor contingency theory - the use of visual stimuli, the learning of visual sensory-motor contingencies and the goal of winning -, but that would not imply that this robot has some kind of conscious visual experience: "Surely someone could accept all that O&N offer, but treat it simply as an account of how certain visual experiences get their contents, rather than as a dissolution of the so-called hard problem of visual qualia." (2001, p. 979).

Martin Kurthen does not understand how O'Regan and Noë can reject qualia, understood as states/occurrences, but maintain the existence of experiences with qualitative character understood as ways of acting. If experience has a qualitative/phenomenal feature "then the gap opens between these features and the ways of acting they are meant to be identified with" (2001, p. 990), that is, "by merely *postulating* an identity of action and visual consciousness, they will not escape the explanatory gap problem, since 'ways of acting' are by no means closer to experiential features than 'internal representations' are." (2001, p. 990). Strictly speaking, the theory of sensorimotor contingencies does not answer the fundamental question: "why should skilled exercise generate phenomenal consciousness at all?". (2001, p. 991). So, "to dismiss qualia in favor of ways of acting, will not suffice to avoid the gap as long as the existence of experiences with qualitative character is affirmed." (2001, p. 991).

Klaus Oberauer agrees with this criticism by saying that the theory proposed by O'Regan and Noë "is in a no better position than any other theory to solve the 'hard problem' of consciousness." (2001, p. 996). He begins his comments by questioning the thesis that qualia, defined as "properties of experiential states or events", is an illusion, after all, by characterizing experiences as "modes of act" or "things we do", even they are not "states", "they certainly are events", that is, "it seems completely reasonable to characterize qualia as features of events going on during perceptual activity." (2001, p. 996). Hence, to say that experience could not be characterized as something static (state) does not imply that the term qualia is devoid of meaning, which can be understood as "a descriptive term that captures the fact that we experience something while we perceive, and that this experience has a certain quality that could be different for different people even if we perceived (in an information-processing sense) the same thing." (2001, p. 996). Like Clark and Toribio, Oberauer thinks that if qualia, as just defined, could be explained from sensorimotor contingencies, there would be a logical necessity for it to occur whenever sensorimotor contingencies were present, what implies that if there were a robot with appropriate sensorimotor contingencies, necessarily this

machine would have a "rich inner life". Oberauer concludes by ironically saying that he stays “agnostic on this, even for a very graceful robot." (2001, p. 996).

According to the criticisms previously summarized, O’Regan and Noë misunderstood the notion of qualia, and consequently the problem of the explanatory gap. Following these criticisms I think that all that O’Regan and Noë accomplished was to establish the essential role of the environment, the body (brain included), and action to the existence and nature of the contents of visual sensations, offering a way of thinking about the existence of all forms and contents of consciousness. But this could be accepted even by those who, like me, believe that the conscious mind is a property of the brain. I think, as I will argue throughout next section, that it is possible intelligibly to address the problem of explanatory gap by considering the embodied and situated brain as the organ of consciousness.

3 Visual consciousness, brain, body, environment and action: addressing the explanatory gap

3.1

By assuming that consciousness is a nonphysical, qualitative, and subjective property of the embodied and situated brain, two problems involving the relationship between brain and consciousness are constitutive of the explanatory gap, namely, that of explaining how the brain causes consciousness – upward causation -, and how consciousness causes brain events – downward causation.

Regarding the first problem, the main difficulty is to explain how the structural and functional complexity of the brain are causally related to the existence of consciousness in general, and, particularly, to specific forms and contents of consciousness. While currently available technologies are contributing significantly to such understanding, deeper knowledge depends on the development of even more sophisticated technologies, which will probably be available in the coming years. If the advance of knowledge of the structure and functioning of the brain related to consciousness will meet limits, this will be due to limits of technological development. If we believe that this gap is insurmountable because of the way our cognition works, so the explanatory gap would not be a unique to that psychophysical relationship, but intrinsic to scientific explanations in general, at least if it is assumed that the scientific solution to a problem consists in identifying spatial and/or temporal correlations between phenomena that precede and/or follow each other regularly, as we learn with David Hume. If the scientific procedure is appropriate for dealing with physical entities/processes, it is also legitimate for handling the relationship between brain and consciousness. As David Chalmers says:

> There is a system of laws that ensures that a given physical configuration will be accompanied by a given experience, just as there are laws that dictate that a given physical object will gravitationally affect others in a certain way. It might be objected that this does not tell us what the *connection* is, or *how* a physical configuration gives rise to experience. But the search for such a

> connection is misguided. Even with fundamental physical laws, we cannot find a "connection" that does the work. Things simply happen in accordance with the law; beyond a certain point, there is no asking "how". As Hume showed, the quest for such ultimate connections is fruitless. If there are indeed such connections, they are entirely mysterious in both the physical and psychophysical cases, so the latter poses no *special* problem here.
> It is notable that Newton's opponents made a similar objection to his theory of gravitation: *How* does one body exert a force on another far away? But the force of the question dissolved over time. We have learned to live with taking certain things as fundamental. (1996, p. 170).

Accepting the legitimacy of this explanatory model, according to which we should not look for an ultimate element that connects events, I present below an outline of a neuroscientific view, according to which visual consciousness is a property of the brain derived from its relationship with the body where it is embodied, and with the environment where brain and body are situated. According to that view, the external physical environment provides not only the objects visually perceived, but also the light, which, reflected by these objects, makes it possible to see them, being thus essential for perceiving colors, shapes, depth, movement, etc. The eye, whose structure is highly complex - pupil, iris, cornea, anterior and posterior chambers, crystalline, and the retina with its sophisticated and complex cellular/molecular constituents, etc.-, is also notoriously indispensable for the existence and characteristics of visual experience, since it allows the capturing of the light reflected by external objects, and from it, other of their properties such as color, form, depth, movement, etc. As is widely known, if the visual organ has some structural and/or functional problem, the quality of the visual experience is affected, as exemplified by cases such as myopia, astigmatism, presbyopia, amblyopia, cataracts, color blindness, total blindness, etc. Besides, it could not be ignored how necessary are specific ocular movements - saccadic and extraocular muscles movements - to make possible the visual consciousness.

What about the role of the brain in visual consciousness? Certainly, it is not possible, and not even necessary, to present in the short space of this paper everything that is already known, and how much is not known about the visual brain. So I will summarize it by saying that from the retina - composed of photoreceptor cells (cones and rods), as well as horizontal, bipolar, amacrine, ganglion, etc. cells - information is transmitted by the nerve and optic structures, such as the lateral geniculate nucleus with its optical radiations, towards the visual cortex, with its several nuclei (V1, V2, etc.) responsible for specific visual functions, and toward areas such as frontal ocular field, superior colliculus, pretectal nuclei, extraocular muscles, etc. The microscopic and macroscopic recording of activities of these structures, and the observation of what happens to the contents of visual consciousness when they are affected by lesions, pharmacological interventions, electromagnetic, etc., has permitted the identification of the function of each one of them, isolated or together, related to the generation of both, the most basic sensory characteristics of visual consciousness, such as color, shape, movement,

spatial location, identification of objects, etc., as well as to the motor processes, also necessary for the generation of those visual contents.

So, I am assuming that this is the direction to be followed by non-reductionist physicalists to explain how the brain, from its interaction with the other parts of the body and with the environment outside the body, produces visual consciousness. In this sense, the *difficult problem* to be solved, the first aspect of the explanatory gap, is a *scientific* problem, which refers to the knowledge of the details of the brain structure and functioning responsible for the several aspects of visual consciousness.

3.2

The other *difficult problem,* the second aspect of the explanatory gap, the downward causation problem, still taking visual consciousness as a paradigmatic example, can be formulated as follows: How does visual consciousness, being a nonphysical, qualitative and subjective property of the brain, cause brain events? I have argued (COELHO, 2017) that this sort of formulation of the problem of mental causation, paradigmatically presented by Jaegwon Kim (1998), is an important part of the problem itself. Intrinsic to that formulation is the idea that consciousness is a nonphysical (immaterial) Cartesian *substance*,[2] or a property of a nonphysical (immaterial) Cartesian *substance*, which would causally affect the brain. Hence the Cartesian problem: How does a nonphysical (immaterial) *substance* act causally on a physical (material) *substance*?

I think that this problem would not arise if visual consciousness was neither considered as a nonphysical *substance*, nor as a property of a nonphysical *substance*, but rather as a nonphysical *property* of the brain, which gives the brain skills that it would not have without visual consciousness. Consciousness in general, and visual consciousness in particular, would be a sort of guide that the brain uses to interact with the body in which it is embodied, and through its body, to interact with the outside world in which brain and body are situated. So the relationship between visual consciousness, brain, body and external world could be summarized as follows: An embodied brain receives, through the eyes, physical stimuli from the external environment from which the brain not only produces the forms and contents of its visual consciousness, but also uses them to guide its embodied movements in the outside environment in which it is situated. The visual consciousness would be for the brain something like the light of a flashlight is to an individual in a dark environment. The flashlight produces the visibility that the individual uses as a guide from which he moves the flashlight creating other

[2] In Article 51 of "Les Principes de la Philosophie", Descartes defines substance as the existing being that does not depend on another being to exist, that is, that is not an attribute of another existing being; on the contrary, it serves as the substrate or support of other existing beings, which would be its attributes: "When we conceive the substance, we conceive only a thing that exists in such a way that it only needs its own to exist." (DESCARTES, p. 594). Article 52 presents the same position, only adding that the substance is a divine creation that does not depend on another divine creation: "to understand what substances are, it suffices only to see that they can exist without the help of anything else created." (DESCARTES, p. 594).

visibilities and so on, in such a way that this allows him to move safely in his environment. Although the individual depends on the flashlight and on the light produced by it to move safely in his environment, it is the individual that moves and guides the flashlight and not the flashlight, let alone the light that moves or guides him.

How could this analogy help us to understand the relationship between the visual consciousness and the brain, considering the conscious mind essential for survival, although not exerting downward causal action? First, let us consider the difference. In the flashlight case there is a relationship between the individual, the flashlight and the light in which the individual, by turning on the flashlight, only indirectly produces the light that he uses, since, strictly speaking, it is the flashlight that produces the light. Considering the light (visibility) as if it were the visual consciousness, the eye as if it were the flashlight and the brain as if it were the individual, the brain and not the eye is what creates visual consciousness (the individual, not the flashlight) from the environmental stimuli captured and transmitted by the eye. Second, let us consider the similarity. Just as the light (visibility) produced by the flashlight is used by the individual as a guide, the visual consciousness produced by the brain is also used by the brain as a guide to act adaptively, through its body, in the outside environment. The actions implemented by the brain upon the body, and throughout the body upon the external environment, from the forms and contents of its visual consciousness, allow it to receive other visual stimuli, which produce other brain events, generating other forms and contents of visual consciousness (new visibilities), which will be used by the brain as guides for new actions in the outside world, and so on.

I believe that the view presented in this section constitutes an intelligible way to address the problem of the explanatory gap, preserving the idea that consciousness has a subjective and qualitative nature produced by the brain from its interaction with its body (including bodily movements) and with the environment outside the body. In this sense, I find it promising to pursue the idea that the brain is the true agent -it is the brain that makes decisions, initiates and supports actions -, which uses the several forms and contents of its conscious mind – memory, belief, intention, volition, emotion, imagination, etc. - as a guide to initiate and maintain actions in the world. The fact that the brain is the agent does not diminish the relevance of the conscious mind, it does not make it an epiphenomenon of the brain, since without the conscious mind the brain would not be, structurally and functionally, what it is - as I have argued in COELHO, 2018) –, and it would not be able to do what it does. I am assuming that the conscious mind is a phenotype created by biological evolution, a property of the brain, which results from its interaction with the body and the external environment, used by the brain as a guide that allows it to act adaptively in the world. I believe that this view is a reasonable path to address the problem of the relationship between the conscious mind and the body, by moving away from the spell of the explanatory gap.

References

CHALMERS, David. *The conscious mind*: In search of a fundamental theory. New York/Oxford: Oxford University Press, 1996.

CLARK, Andy; TORIBIO, Josefa. Sensorimotor chauvinism? *Behavioral and Brain Sciences*, v. 24, n. 5, Oct. 2001, pp. 979 - 980.

COELHO, Jonas Gonçalves. A double face view on mind-brain relationship: the problem of mental causation. *Trans/Form/Ação*, v. 40, n. 3, p, 197-220, 2017.

_____. Abordagem dupla face da relação mente consciente e cérebro: a visão como exemplo paradigmático. In: TOLEDO, Gustavo Leal; GOUVEA, Rodrigo A.; ALVES, Marco Aurélio S. (Orgs.). *Debates contemporâneos em filosofia da mente*. São Paulo: FiloCzar, 2018. p. 131-150.

DEGENAAR, Jan; O'REGAN, J. Kevin. Sensorimotor theory of consciousness. *Scholarpedia*, v. 10, n. 5, 2015, p. 4952. Available at: http://www.scholarpedia.org/article/Sensorimotor_theory_of_consciousness.

DESCARTES, René. Les Principes de la philosophie. In: *Oeuvres et Lettres*. Paris: Librairie Gallimard, 1952.

HURLEY, Susan; NOË, Alva. Neural plasticity and consciousness. *Biology and Philosophy*, v. 18, n. 1, 2003, pp. 131-168.

KIM, Jaegwon. *Mind in a physical world*: an essay on the mind-body problem and mental causation. Cambridge: MIT Press, 1998.

KURTHEN, Martin. Consciousness as action: the eliminativist sirens are calling. *Behavioral and Brain Sciences*, v. 24, Oct. 2001, pp. 990-991.

LEVINE, Joseph. Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly*, v. 64, n. 4, 1983, pp. 354-361.

NAGEL, Thomas. What is it like to be a bat? *Philosophical Review*, v. 83, n. 4, 1974, pp. 435-450.

OBERAUER, Klaus. The explanatory gap is still there. *Behavioral and Brain Sciences*, v. 24, n. 5, Oct. 2001, p. 996-997.

O'REGAN, J. Kevin.; NOë, Alva. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, v. 24, n. 5, Oct. 2001, pp. 939-1031.

Recebido em: 21-02-2019

Aceito para publicação em: 28-08-19

(RE-)INTERPRETANDO "THOUGHT & TALK": DONALD DAVIDSON ACERCA DAS MENTES ANIMAIS*

(RE-)ASSESSING "THOUGHT & TALK": DONALD DAVIDSON ON ANIMAL MINDS

DIANA COUTO[1]
Universitat de Barcelona – Espanha
dpcouto@ub.edu

RESUMO: De acordo com a interpretação mais comum na literatura filosófica, Donald Davidson — que celebremente afirmou que "uma criatura não pode ter pensamentos a menos que tenha uma linguagem" — nega que criaturas não linguísticas são criaturas pensantes. No entanto, neste artigo argumento que esta interpretação é errada. Analisando atentamente os argumentos de Davidson, procuro mostrar que ele não está a argumentar que criaturas não linguísticas não podem possuir pensamentos; em vez disso, defendo que ele está simplesmente a afirmar que na ausência de linguagem as atribuições de pensamento não podem ser vistas como mais do que meras ficções úteis que permitem explicar com êxito o comportamento daquelas criaturas.

PALAVRAS-CHAVE: Crença *de re/de dicto*. Linguagem. Triangulação. Indeterminação. Interpretação.

ABSTRACT: *It is commonly held in the philosophical literature that Donald Davidson, who famously claimed that "a creature cannot have thoughts unless it has a language", denies that languageless creatures are thinking creatures. In this paper, however, I argue that this interpretation is arguably mistaken. By looking more closely at Davidson's arguments, I attempt to show that he is not arguing that dumb creatures cannot have thoughts, but rather that in the absence of language thought-attributions cannot count as more than useful fictions that successfully account for the behavior of those creatures.*

KEYWORDS: *De re/de dicto belief. Language. Triangulation. Indeterminacy. Interpretation.*

Do animals have beliefs? To paraphrase my young son:
"A little bit they do. And a little bit they don't"
(STICH, 1979, p. 28).

*A elaboração deste artigo contou com o financiamento da bolsa de doutoramento SFRH/BD/129112/2017 atribuída pela Fundação para a Ciência e a Tecnologia (Portugal).

[1] Doutoranda em Filosofia. Membro do LOGOS - Research Group in Analytic Philosophy; BIAP - Barcelona Institute of Philosophy; MLAG - Mind, Language, and Action Group, Instituto de Filosofia da Universidade do Porto.

Introdução

Pouco mistério existe em relação ao conteúdo da nossa mente e da mente de outros. Se tivermos curiosidade em saber quais as inclinações políticas ou desportivas de alguém, ou pretendemos saber por que motivo o nosso vizinho atravessou a rua quando o semáforo ficou vermelho, podemos, entre muitas coisas, perguntar, ler o que escreveram, observar a sua conduta. Com base nestas descobertas podemos explicar, e quem sabe até prever, muitos dos seus comportamentos futuros. Se por um lado é o caso — como eu penso que é — que cada um de nós tem um acesso de algum modo privilegiado aos seus conteúdos mentais sem que para isso tenha de interpretar as suas próprias enunciações ou observar o seu próprio comportamento, então talvez possamos dizer, por outro lado, que a presença da linguagem é o que permite atenuar o segredo acerca dos conteúdos mentais daqueles com quem nos podemos comunicar.

Imagino que muitos se sentirão atraídos pela crença de que, tal como muitos humanos adultos linguisticamente competentes possuem pensamentos, pelo menos algumas criaturas não linguísticas como bebés humanos, símios e cães também os possuem. Em última análise, quem não se sente tentado a concordar que o cão de Norman Malcolm (1972-3), que ladra insistentemente para *aquela* árvore no jardim, acredita que o gato que persegue está escondido entre os seus ramos? Independentemente de se nos inclinamos a responder de forma afirmativa ou negativa a esta questão, enfrentamos uma disputa muito mais exigente acerca dos fundamentos que possuímos para justificar essa resposta. É este o único problema que, à luz de Donald Davidson, pretendo tratar aqui. Suspeito que assumir uma posição neste debate não seja tão fácil como à primeira vista poderá parecer.

O assunto aqui em causa tem uma longa história. Afinal, qual é a relação entre a linguagem e o pensamento? Será o pensamento dependente da linguagem? Ao longo dos tempos foram dadas três diferentes respostas a esta questão:[2]

> *(i) Lingualismo*: Afirma que o pensamento depende da linguagem, pelo que se certas criaturas não possuem linguagem, então essas criaturas não são criaturas pensantes.
>
> *(ii) Mentalismo*: Sustenta que o pensamento não depende da linguagem, de maneira que criaturas não linguísticas podem possuir pensamentos. Devido à diferença nos inputs e arquitetura percetivos em criaturas linguísticas[3] e criaturas não linguísticas, os pensamentos possuídos pelas últimas diferem *em grau* daqueles possuídos pelas primeiras.
>
> *(iii) Posição intermédia entre (i) e (ii)*: mantém que criaturas não linguísticas podem possuir pensamentos de um género mais simples do que os pensamentos possuídos por criaturas linguísticas, nomeadamente

[2] Sigo Hans-Johann Glock (1999) e Christian Barth (2011, p. 01) aqui.
[3] Por criatura linguística entende-se uma criatura capaz de manifestar respostas verbais.

pensamentos que podem ser expressos através de comportamentos não verbais.[4]

Antes de prosseguir, é conveniente esclarecer o que aqui se pretende dizer com "dependência". Ora, a dependência aqui em causa deve ser entendida no sentido de *prioridade explicativa* ou *conceptual* que um dos termos — neste caso, "pensamento" ou "linguagem" — poderá ter sobre o outro.[5] Por outras palavras, o que se tem em mente quando se fala de prioridade explicativa é a eventual possibilidade de um dos termos poder ser elucidado à luz do outro. Com respeito a (i), afirmar que o pensamento depende da linguagem e que, por conseguinte, pode ser elucidado à luz desta deve ser entendido da seguinte maneira: pensamentos são tipicamente expressos por meio da linguagem, pelo que se uma dada criatura não possui linguagem, então essa criatura não possuirá os pensamentos que são tipicamente expressos através dela. De acordo com esta interpretação, a ausência de linguagem é um indício da ausência de pensamento, pois o último só pode ser explicado por referência à primeira. Inversamente, quanto a (ii), que sustenta que o pensamento não depende da linguagem, esta dependência deve ser entendida da forma que se segue: há criaturas indisputavelmente diferentes entre si, quer em complexidade biológica, quer nas suas capacidades percetivas e cognitivas. Ainda que uma criatura não possua linguagem, essa criatura poderá possuir outro tipo de pensamentos que não têm necessariamente de ser expressos por meio de linguagem. Nesse caso, a ausência de linguagem não constitui nenhum indício da ausência de pensamento. Note-se que a diferença crucial entre (i) e (ii) reside no facto de (i) eliminar qualquer cenário em que uma criatura não linguística possua pensamentos, ao passo que (ii) deixa aberta essa possibilidade.

De acordo com a interpretação generalizada na literatura filosófica, Davidson é um dos mais célebres defensores de (i). A esta interpretação chamarei *Leitura Forte* (daqui para a frente: *LF*). A sua afirmação de que "uma criatura não pode ter pensamentos a menos que tenha uma linguagem" (1982, p. 100) despoletou o debate entre filósofos, etólogos e donos de animais de estimação, tendo vindo a servir de mote para várias críticas provenientes de ambas as partes (veja-se, por exemplo, MIGUENS, 2006; MIGUENS & PINTO, 2018). Muitas destas críticas apoiam-se em investigações empíricas e colocam em cheque pelo menos uma das premissas dos alegados argumentos de Davidson a favor de (i). Apesar de Davidson nunca ter simpatizado com a ideia de que estudos empíricos são relevantes para a análise filosófica e, por conseguinte, ter levado a cabo uma investigação filosófica puramente conceptual e *a priori*, não pretendo aqui atacar o fundamento destas críticas, muito menos acusar os críticos de Davidson de estarem a mudar o tema em discussão. Neste artigo, vou igualmente recorrer a alguns destes estudos empíricos de modo a argumentar que eles mesmos podem servir de suporte à própria conclusão que pretendem rebater. Por outras palavras,

[4] A minha atenção neste artigo dirige-se exclusivamente às posições (i) e (ii), pelo que nada terei a dizer acerca de (iii).
[5] Retomo a expressão "prioridade explicativa ou conceptual" de Davidson (1975, pp. 155-156).

vou argumentar que estes argumentos falham em *demonstrar* a presença de pensamento na ausência de linguagem. A minha linha de raciocínio levar-me-á a defender uma *Leitura Débil* (daqui para a frente: *LD*) da posição de Davidson em relação às mentes animais segundo a qual Davidson não está comprometido com uma leitura estrita de (i). Ao invés, as suas afirmações são melhor entendidas como a defesa de uma quarta via, cética quanto à posse de pensamento na ausência de linguagem:

> *(i) Ceticismo*: devido à incapacidade de certas criaturas manifestarem respostas verbais, não estamos em posição nem de afirmar, nem de negar, que estas possuam pensamentos.

Para este fim, em §2 esclareço o que entendo por *LF* e as consequências que lhe estão associadas, nomeadamente a aceitação de um certo verificacionismo relativamente ao que pode ser conhecido. Em §3 apresento algumas abordagens ao tópico das mentes animais. Em §4 discuto os argumentos de Davidson que supostamente sustentam a *LF* e defendo que os proponentes desta leitura falham em identificar o alvo dos mesmos. Finalmente, em §5 argumento que Davidson defende um ceticismo quanto à possibilidade de criaturas não linguísticas possuírem pensamentos que assenta na drástica indeterminação explicativa dos comportamentos destas.

1 *Leitura Forte* e a premissa verificacionista

A dificuldade que surge na hora de responder à pergunta "Criaturas não linguísticas possuem crenças?" resulta da tensão que surge das duas suposições seguintes:[6,7]

(1) Criaturas não linguísticas possuem crenças.[8]

(2) O conteúdo das crenças de criaturas não linguísticas não pode ser devidamente caracterizado.

[6] Sigo Dale Jamieson (2009) aqui.
[7] Usarei os termos "pensamento", "crença" e "racionalidade" num sentido davidsoniano, i.e., como atitudes proposicionais: atitudes que um sujeito toma perante uma dada proposição.
[8] Alguns podem sentir-se inclinados a acreditar que esta questão está relacionada com a simplicidade/complexidade biológica do organismo em causa e que, precisamente devido a essa simplicidade/complexidade, nem todas as criaturas não linguísticas pensam, mas pelo menos algumas o fazem. Há certamente uma diferença significativa entre organismos unicelulares e, por exemplo, símios. Contudo, dado que o que me ocupa neste artigo concerne à *justificação* das atribuições de crenças, acredito que esta diferença, ainda que não seja irrelevante para a discussão em torno das mentes animais, não é crucial nem afetará os meus argumentos uma vez que os mesmos não se prendem a considerações biológicas e aplicam-se igualmente a sistemas não biológicos. No máximo, poderíamos falar em graus de justificação na atribuição de crenças: atendendo à complexidade biológica de certas criaturas, estamos *mais justificados* a atribuir-lhes crenças do que a outras criaturas. No entanto, esta não é a linha de argumentação que pretendo seguir aqui.

Para compreender o problema em causa é necessário, em primeiro lugar, distinguir (1) da *utilidade pragmática das atribuições de crença* (daqui para a frente: *UPAC*) (FELLOWS, 2000, p. 587):

> (*UPAC*) Graças ao êxito na explicação e previsão do comportamento de criaturas não linguísticas através da atribuição de crenças, esta atribuição revela-se útil e está pragmaticamente justificada.

A diferença entre (1) e a (*UPAC*) reside no seguinte: ao passo que (1) parte da ausência de linguagem por parte das criaturas para a estipulação de que elas não possuem, literalmente, crenças, a (*UPAC*) afirma, somente, que as atribuições de crença a criaturas não linguísticas estão *pragmaticamente justificadas* devido ao êxito explicativo e preditivo dos seus comportamentos. Deste modo, a (*UPAC*) nada tem a dizer acerca de se a utilidade ou justificação pragmática da atribuição de crenças constitui com efeito uma demonstração da presença destas atitudes nas criaturas em questão.

A *LF* toma Davidson como um defensor do chamado lingualismo introduzido em §1, ao qual está associado um certo tipo de verificacionismo dificilmente defensável que se apresenta como o seguinte *reductio ad absurdum*:[9]

(P1) Muitas criaturas não linguísticas pensam.

(P2) Pensar implica pensar sobre algo em particular.

(P3) O que criaturas não linguísticas pensam não pode ser caracterizado de forma fiável.

(P4) Não podemos verificar a verdade de (P1).

(P5) Devemos acreditar somente no que podemos verificar. (Premissa verificacionista)

(P6) Não devemos acreditar em (P1).

(P1) e (P3) correspondem, respetivamente, às suposições (1) e (2) mencionadas anteriormente. (P1) é a premissa para a qual pretendemos encontrar uma justificação adequada, i.e., uma justificação que *demonstre* a posse de crenças e que vá mais além do que uma justificação *pragmática* do comportamento de criaturas-não linguísticas. Procurarei mostrar, mais adiante, que Davidson não rejeita absolutamente, nem aceita uma leitura estrita, de (P1). (P3), por sua vez, é aceite por Davidson que apresenta três argumentos relacionados para apoiar a sua defesa. Estes argumentos serão discutidos em §4.1, §4.2 e §4.3. (P4) segue-se de (P3) e é também aceite por Davidson. (P2), ao contrário de (P5) e (P6), não parece problemática. O que há a saber é se Davidson, por aceitar (P3) e (P4), está comprometido com a aceitação de (P5) e (P6). A minha resposta é negativa. De acordo com a *LD* que proponho, Davidson aceita uma versão mais fraca de (P5) e (P6), a saber,

[9] Sigo a reconstrução de Jamison (2009, p. 19).

(P5*) Podemos demonstrar somente o que podemos verificar.

(P6*) Não temos fundamentos para demonstrar (P1).

Como irei mostrar mais à frente, a aceitação de (P5*) e (P6*) por parte de Davidson não entra em conflito com o seu ceticismo quanto a (P1).

O verificacionismo aqui em causa não é, naturalmente, o verificacionismo segundo o qual o significado deve ser entendido em termos das suas condições de verificação. Como é sabido, Davidson defende uma teoria verocondicional do significado (veja-se WILLIAMSON, 2004). O verificacionismo aqui aludido caracteriza-se pela forte suposição acerca daquilo que pode ser conhecido; nas palavras de Kathrin Glüer, trata-se da "alegação de que um certo conjunto de factos é tal que pode, em princípio, ser conhecido" (2011, p. 133). Este verificacionismo assenta na confusão entre questões de índole epistemológica e questões de índole ontológica. Negar algo pelo facto de não ser possível conhecer ou verificar a sua verdade seria o resultado desta confusão. Segundo a *LF*, Davidson rejeita (P1) apoiando-se para isso em (P3). Se isto for correto, então acusação de verificacionismo é-lhe bem aplicada. Contudo, embora a suposição subjacente à *LF* não seja completamente descabida, ela não tem de ser a última palavra e, a meu ver, o que Davidson nos diz é, apenas, que a atribuição de crenças é útil e está pragmaticamente justificada, mas não constitui um indício forte, nem a favor nem contra, a verdade de (P1). Esta utilidade, por sua vez, embora possa eventualmente servir para atenuar (P3), continua a deixar o comportamento de criaturas não linguísticas indeterminado de um modo tão drástico como inaceitável, tornando qualquer atribuição de crenças infundada. Por outras palavras, Davidson aceita (P3) e a (*UPAC*), mas mantém-se cético quanto à verdade, ou falsidade, de (P1).

Um dos aspetos primordiais a prestar atenção é que Davidson não se assume um defensor do lingualismo, nem do mentalismo, nem muito menos de uma posição intermédia entre ambos. Para ele, linguagem e pensamento são interdependentes, o que significa que não há prioridade explicativa de um sobre o outro.[10] À primeira vista, podemos suspeitar que esta interdependência pontua a favor da *LF* na seguinte medida: a fim de dar sentido ao comportamento verbal de um falante, o intérprete deve atribuir-lhe, simultaneamente, crenças e significados. Na ausência de comportamento verbal não haverá nada a que atribuir significado e, consequentemente, não será possível justificar a atribuição de crença. Perante isto, não estaríamos em posição de considerar a criatura em questão uma criatura racional. As palavras de Davidson parecem não deixar dúvidas quanto a isto:

> Sabemos como são estados mentais, e como são corretamente identificados; são apenas aqueles conteúdos que podem ser descobertos de maneiras bem conhecidas. Se outras pessoas ou criaturas estão em estados que não podem ser descobertos através

[10] Davidson aceita a circularidade entre crenças-significados alegando que não é possível conhecer os significados das enunciações de um falante sem conhecer previamente as suas crenças e, ao mesmo tempo, não é possível inferir as suas crenças se não se é capaz de compreender as suas enunciações. Veja-se Davidson (1973, 1974, 1975).

> destes métodos, não se trata de que os nossos métodos falham, mas que aqueles estados não são corretamente chamados estados mentais — não são crenças, desejos, anseios ou intenções. (DAVIDSON, 1988, p. 40).

Para compreender a posição de Davidson relativamente às mentes animais — ou, de modo mais geral, em relação ao mental —, bem como para compreender o sentido desta interdependência entre linguagem e pensamento é essencial conhecer o propósito do seu projeto filosófico: tornar inteligível — *explicar* — um determinado comportamento, verbal ou não verbal (veja-se MIGUENS, 2004). A interdependência linguagem-pensamento diz-nos que nos casos em que a criatura é capaz de enunciar respostas verbais, a atribuição de crença não está apenas pragmaticamente justificada, como também constitui uma demonstração da sua posse. No caso inverso, isto não se verifica. Uma vez mais, o motivo remete-nos para a indeterminação da interpretação. Embora a indeterminação também ocorra nos casos de interpretação linguística, na ausência de respostas verbais por parte de uma criatura esta torna-se drasticamente radical. Esta radicalidade torna impossível identificar na criatura em questão características que Davidson toma como necessárias à posse de pensamento, nomeadamente a capacidade de distinguir entre aquilo que é subjetivamente verdadeiro e objetivamente verdadeiro, entre crença e realidade. Retomarei este assunto mais à frente.

2 Algumas abordagens

Muitos teóricos aceitam que (*UPAC*) é suficiente para demonstrar (P1) (por exemplo, DENNETT, 1998). Estes apoiam-se no facto de certas criaturas não linguísticas exibirem um *comportamento dirigido a fins* que bastaria, à primeira vista, para rejeitar – ou, pelo menos atenuar – (P3), somando pontos a favor de (P1). À primeira vista, esta posição parece bem suportada por algumas investigações empíricas levadas a cabo na psicologia cognitiva que mostram que certas criaturas não linguísticas não só manifestam comportamentos dirigidos a fins, como também são capazes de exercer comportamentos de ordem mais complexa – por exemplo, comportamento de engano – que seriam em si suficientes para justificar a atribuição de crenças e outros estados mentais (TOMASELLO, 1997; TOMASSELO, CALL & MARLER, 1980).

A nível de senso comum, o êxito na explicação do comportamento de criaturas não linguísticas por meio da atribuição de crenças parece constituir um indício a favor de (P1). Em última análise, é este êxito explicativo que nos faz concordar com Malcolm na sua muito comentada passagem que cito em seguida:

> Suponha-se que o nosso cão está a perseguir o gato do vizinho. O gato do vizinho corre em direção ao carvalho, mas de repente, no último momento, desvia-se e desaparece num ácer próximo. O cão não vê esta manobra e, ao chegar ao carvalho, levanta-se sobre as patas traseiras, com as patas no tronco como se tentasse escalá-lo, e ladra ansiosamente para os ramos acima. Nós, que observamos

> todo este episódio de uma janela, dizemos "Ele pensa que o gato subiu aquele carvalho". (MALCOLM, 1972-3, p. 13).

Estaríamos justificados a atribuir ao cão de Malcolm a crença de que o gato trepou à árvore? Davidson (1982, pp. 101-102) não só diria que sim, como também diria que carecemos de uma melhor explicação para o seu comportamento. Com isto, ele estaria a aceitar a (*UPAC*). Porém, é importante ter em conta que a sua aceitação de (*UPAC*) não implica nem a negação de (P3), nem a aceitação estrita, ou negação absoluta, de (P1). A aceitação de (*UPAC*) por Davidson diz-nos que estamos tão justificados a atribuir ao cão de Malcolm a crença de que o gato trepou à árvore, como a atribuir a um míssil térmico o desejo de afundar o Bismark e a crença de que, ao seguir uma determinada trajetória, afundará o Bismark. Naturalmente, Davidson reconhece que o comportamento de criaturas não linguísticas é mais semelhante, quer em complexidade, quer em imprevisibilidade, ao comportamento de humanos do que o comportamento de mísseis térmicos, da mesma forma que reconhece que, ao contrário do comportamento de criaturas não linguísticas, possuímos uma explicação mais básica para o comportamento do míssil. Com esta analogia Davidson pretende apenas chamar a nossa atenção para o facto de que quando atribuímos crenças tendo por base unicamente a observação de comportamento não verbal das criaturas podermos estar a recorrer a um vocabulário antropológico e, portanto, "demasiado sofisticado" para o comportamento que pretendemos explicar.

Ao aceitar a distinção entre (P1) e (*UPAC*), Davidson está a aceitar que o comportamento não verbal serve como justificação pragmática das atribuições de crença, mas não constitui uma demonstração da sua posse por parte das criaturas em causa. Mais à frente irei mostrar que o impulso de Davidson para defender esta posição apoia-se em considerações normativas: ao atribuir crenças a criaturas não linguísticas estamos a tratá-las *como se* elas estivessem a, ou *fossem capazes de*, *agir por uma razão*, o que não implica que elas estejam, efetivamente, a *agir por uma razão*. Se as criaturas não forem capazes de verbalizar as razões que as levaram a exercer um determinado comportamento, a explicação do mesmo ver-se-á seriamente indeterminada.

Stephen Stich (1979) faz afirmações que apontam na mesma direção das de Davidson. Ele declara que não temos como determinar o conteúdo das crenças de criaturas não linguísticas e, em consequência, aceitar a (*UPAC*) não implica nem rejeitar (P3), nem constitui uma demonstração de (P1). Se insistirmos em atribuir crenças a criaturas não linguísticas com base no seu comportamento, então poderemos estar a usar um vocabulário demasiado sofisticado para o fenómeno a explicar. Nos termos de Glock (1999), o *explanans* excede o *explanandum*.

Embora aparentemente inócua, a (*UPAC*) está longe de ser consensual entre os filósofos. Edward J. Lowe (2004, pp. 178-179), por exemplo, rejeita a (*UPAC*) alegando que incorre em petição de princípio. Por um lado, se fundamentamos a (*UPAC*) no facto de certas criaturas não linguísticas manifestarem uma conduta inteligente — i.e., uma conduta semelhante à de humanos, aparentemente ordenada, não arbitrária e que parece dirigir-se à satisfação de certos propósitos

—, e se entendemos que "conduta inteligente" envolve a posse de crenças ou outras atitudes proposicionais, então a nossa justificação é circular. Por outro lado, se por "conduta inteligente" entendemos somente uma conduta que se adapta às necessidades da criatura tendo em conta as circunstâncias do meio onde está inserida, a nossa incógnita inicial permanece: não sabemos se (P1) é verdadeira nem como demonstrá-la, visto que é possível dar uma explicação mais básica do mesmo comportamento entendendo-o, por exemplo, como uma resposta a determinados estímulos perceptivos sem que para isso tenhamos de recorrer a vocabulário intencional. Lowe defende que as atribuições de crença demonstram (P1) somente se a criatura cujo comportamento se pretende explicar for capaz de abdicar da satisfação dos seus desejos imediatos tendo em vista a satisfação de desejos futuros como consequência da sua ação presente. Apoiando-se em alguns estudos empíricos levados a cabo por Kohler (1959), Cheyney & Seyfarth (1990) e Heyer & Dickinson (1990), ele argumenta que apenas humanos adultos são capazes de efetuar esta dissociação. Deste modo, Lowe aceita (P3), e rejeita quer (P1) quer a (*UPAC*).

Um dos pontos mais sensíveis de discórdia entre filósofos e psicólogos cognitivos está em saber se a (*UPAC*) é suficiente para rejeitar (P3) e demonstrar (P1). A legitimidade em associar a (*UPAC*) à demonstração de (P1) é bastante discutível. A psicologia cognitiva procura, através da observação do comportamento de certas criaturas não verbais, inferir estados mentais. A filosofia, por sua vez, entra em cena ao procurar clarificar a relação comportamento-mente mediante o estabelecimento de critérios que determinam quando, e em que circunstâncias, estas atribuições estão ou não justificadas. É este ponto de contato que tem vindo a potenciar, ao longo dos anos, o diálogo entre filósofos e cientistas.

É curioso ver que grande parte destes estudos empíricos são o exemplo claro de que o comportamento de criaturas não linguísticas não é capaz, de um ponto de vista conceptual, demonstrar (P1). Isto deve-se sobretudo ao facto de as atribuições de (pelo menos certas) crenças requererem linguagem para a sua manifestação (DRECKMANN, 1999; GLOCK, 1999). Em última análise, foi esta convicção que motivou Davidson a deixar de lado os estudos empíricos quando falamos de pensamento e linguagem, e a defender que o debate em torno das mentes animais "não é inteiramente empírico, pois há a questão filosófica de qual evidência é relevante para decidir quando uma criatura tem uma atitude proposicional" (DAVIDSON, 1982, p. 95).

A dificuldade em estabelecer os critérios para demonstrar (P1) levou alguns filósofos, que aceitam a (*UPAC*), a considerar que estas atribuições não devem ser vistas como algo mais do que uma estratégia útil que nos permite tornar inteligíveis determinados comportamentos. Neste sentido, estas atribuições tratam-se apenas de uma *descrição* de um comportamento em vez de uma *hipótese explicativa* genuína (GLOCK, 1999, pp. 07-08). Ao contrário das descrições, as explicações requerem que sejam enunciadas as razões que estiveram na origem do seu comportamento. Isto é algo que Davidson aceitaria: só a enunciação de respostas verbais por parte da criatura permite distinguir descrições de explicações; em

última análise, só a presença de linguagem permite, aos olhos de Davidson, associar legitimamente a (*UPAC*) à demonstração de (P1).

3 Argumentos a favor da *Leitura forte*

De acordo com a *LF*, Davidson avança três argumentos interrelacionados que têm como objetivo rejeitar (P1): a restrição da intensionalidade, a restrição holística e a restrição do conceito de crença. Seguidamente, vou analisar cada um destes argumentos e mostrar que a *LF* falha claramente a compreender o alvo de cada um deles: o objetivo de Davidson nunca foi rejeitar (P1), mas *somente* defender (P3) e argumentar que a (*UPAC*) não é em si suficiente para demonstrar (P1).

3.1 Intensionalidade

Seja *S* uma criatura arbitrária e *P* uma crença arbitrária. A restrição da intensionalidade pode ser apresentada nos seguintes moldes:

(*RI*) *S* possui *P* somente se *P* figura em contextos intensionais.

Uma característica importante das atribuições de crença reside no facto de estas gerarem contextos intensionais, i.e., contextos nos quais a substituição de termos correferentes na frase que expressa o conteúdo da crença poderá alterar o valor de verdade da mesma. Por exemplo, Óscar pode acreditar em (F1) e (F2) e, ainda assim, não acreditar em (F3), sendo (F1), (F2) e (F3), respetivamente:

(F1) Walter Scott é Walter Scott.

(F2) Walter Scott morreu.

(F3) O autor de Waverly morreu. (Sendo Walter Scott o autor de Waverly.)

Ora, Davidson declara que na ausência de linguagem a substituição de termos correferentes nas frases de crença pode levar-nos a atribuições "absurdas", i.e., a atribuições às quais não saberíamos como atribuir sentido. Vejamos as seguintes frases:

(F4) Fido acredita que o gato trepou àquela árvore do jardim.

(F5) Fido acredita que o gato trepou à árvore mais antiga do bairro.

Suponhamos que o gato trepou à árvore do jardim e que essa árvore é também a mais antiga do bairro. Ora, se nos parece que estamos justificados a fazer a atribuição expressa por (F4) a Fido na medida em que explica de modo apropriado ao seu comportamento, então estamos igualmente justificados a fazer a atribuição expressa por (F5). Mas, pergunta Davidson, será que estamos justificados a fazer a atribuição expressa por (F5) ou sequer qualquer atribuição?

Bernard Williams expõe um exemplo semelhante. Imaginemos um cão cujo dono é o Presidente dos EUA. O Presidente entra na Casa Branca, o cão acorda e abana a cauda. Pela observação do seu comportamento, podemos dizer que o cão está feliz por ver o seu dono. Mas será que podemos dizer que o cão está feliz por ver o Presidente dos EUA? A resposta de Williams é a seguinte:

> Se o dono deste cão fosse o presidente dos Estados Unidos, dificilmente diríamos que o cão considerara esse indivíduo o Presidente dos Estados Unidos. Será assim porque é mais plausível dizer que o cão tem o conceito 'dono' que o conceito 'Presidente dos Estados Unidos'? Por quê? O conceito 'dono' é um conceito que incorpora tanto conhecimento detalhado sobre convenções humanas, sociedade e assim por diante quanto o conceito 'Presidente dos Estados Unidos'. Parece haver tanto convencionalismo ou artificialidade em atribuir a um cão o conceito 'dono' quanto em atribuir-lhe o conceito 'Presidente dos Estados Unidos'. Então, por que estamos mais satisfeitos em dizer que um cão toma um certo indivíduo como sendo o seu mestre do que dizer que ele toma um certo um certo indivíduo como sendo o Presidente dos Estados Unidos? Creio que a resposta a isto tem algo a ver não com o facto de o cão realmente ter um conceito efetivo 'dono', o que seria uma noção absurda, mas com o facto de muito do comportamento do cão ser com efeito condicionado por situações que envolvem alguém sendo o seu dono, ao passo que muito pouco do comportamento do cão é condicionado por situações que envolvem essencialmente alguém como sendo o Presidente dos Estados Unidos. Ou seja, o conceito 'dono' entra em nossa descrição do reconhecimento ou quase-pensamento ou crença do cão, porque esse é um conceito que queremos usar ao explicar grande parte do comportamento do cão. É algo nessas linhas, creio, que justificará a introdução de certos conceitos nos quase-pensamentos ou crenças de um animal, e a recusa em introduzir outros conceitos nos quase-pensamentos ou crenças de um animal. No caso de seres humanos, no entanto, a situação não é assim, porque temos outros testes para saber quais conceitos o ser humano de facto tem. (WILLIAMS, 1973, p. 139).

Retomando o nosso exemplo de Fido, a questão que se impõe é a seguinte: se não sabemos como justificar a atribuição expressa em (F5), o que nos leva a tomar (F4) como mais justificada?

A fim de contornar a opacidade semântica, poder-se-ia parafrasear (F4) e (F5), por exemplo, da seguinte maneira:

(F4*) Fido acredita, relativamente àquela árvore do jardim, que o gato a trepou.

(F5*) A árvore mais antiga do bairro é aquela a que Fido pensa que o gato trepou.

(F4*) e (F5*) diferem de (F4) e (F5) na medida em que anulam a opacidade semântica dos contextos intensionais, sugerindo que a posição que "àquela árvore/a árvore" ocupa em (F4) e (F5) é transparente ou extensional. Deste modo,

através das atribuições *de re* expressas nas frases (F4*) e (F5*) estaríamos a aceitar que existe uma possível descrição do objeto da crença — neste caso, a árvore — que Fido aceitaria sem impor a necessidade de conhecer essa mesma descrição. A vantagem das atribuições *de re* de (F4*) e (F5*) face às atribuições *de dicto* expressas em (F4) e (F5) reside no facto de, através delas, ser possível explicar o comportamento de Fido fazendo referência a um determinado objeto sem impor que Fido tenha de ser capaz de categorizar esse objeto (por exemplo, saber que a árvore do jardim é um carvalho e que um carvalho é diferente de um sobreiro).

Contudo, esta estratégia não nos deixa ir muito longe, pois embora nos permita identificar um objeto que poderia também ser identificado por Fido mediante uma hipotética descrição, a construção *de re* "*relativamente àquela árvore*", expressa em (F4*), requer, no conteúdo proposicional da cláusula-que, uma referência anafórica ao objeto em questão que Fido não seria capaz de reconhecer e distinguir de outros objetos. Davidson afirma que se insistirmos que Fido seria capaz de reconhecer e de distinguir árvores sob uma determinada descrição, então teremos de atribuir a Fido o conceito de árvore. Porém, se Fido tem o conceito de árvore, então ele deverá ter vários conceitos relacionados com este. O problema com que nos deparamos aqui, como veremos na secção seguinte, é que não sabemos como dar sentido a este conjunto de atribuições interrelacionadas (DAVIDSON, 1982, pp. 97-98; GLOCK, 2003, pp. 272-273; 1999, §4).

Com a (*RI*), Davidson (1982, pp. 97-98; 1975, pp. 163-164) diz-nos que as nossas atribuições *de dicto* na ausência de linguagem não constituem uma demonstração de (P1): "[...] na ausência de linguagem não podemos fazer as distinções finas entre pensamentos que são essenciais às explicações que podemos convictamente prover" (DAVIDSON, 1975, p. 163). É por este motivo que, em última análise, ele argumenta que as atribuições de crença a criaturas não linguísticas não devem ser entendidas num sentido que vá mais além da (*UPAC*). Mas estará com isto a rejeitar (P1)? A resposta é negativa:

> Até agora, tenho assinalado a duvidosa aplicabilidade do teste da intencionalidade no que diz respeito a animais mudos, e o requerimento de haver um rico conjunto de crenças gerais (e verdadeiras), se o pensamento está presente. Estas considerações apontam à linguagem, *mas não equivalem a uma demonstração de que a linguagem é necessária ao pensamento*. Com efeito, o que estas considerações sugerem é apenas que provavelmente não pode haver muito pensamento sem linguagem. (DAVIDSON, 1982, p. 101 – itálico acrescentado).

Com isto, vemos que a (*RI*) suporta (P3) mas não pretende, nem implica, rejeitar (P1).

3.2 Holismo

A restrição holística estabelece que uma crença *P* não pode existir independentemente de outras crenças com as quais *P* mantém relações de coerência lógica. Esta restrição pode ser formulada nos seguintes moldes:

> (*RHOL*) *S* possui *P* somente se *S* possui um conjunto de crenças *N* com o qual *P* é coerente.

A (*RHOL*) incide sobretudo na lógica e identidade da crença. Para Davidson (1982, p. 99), *S* não pode possuir *P* se não possui um conjunto de crenças *N* que suportam e identificam o conteúdo de *P.* O que isto significa é que *S* não poderá possuir *P* se desconhecer que *P* implica, ou é implicada por, pelo menos algumas crenças de *N*.

Imagine-se que *S* acredita que *P*. Considere-se, agora, que *P* implica *Q*. Porém, suponha-se que *S* acredita que *P*, mas não acredita que *Q*. De acordo com esta suposição, *S* seria em parte ignorante do conteúdo da sua crença *P.* Segundo Glock (1999), seguem-se daqui dois problemas. Em primeiro lugar, esta ignorância parece incompatível com a autoridade de *S* perante os seus próprios conteúdos mentais. Em segundo lugar, o pensamento de *S* não poderá ser o mesmo que o pensamento de uma criatura *Z* que acredita que *P* e, ao mesmo tempo, acredita que *Q*. Os pensamentos de *S* e *Z* seriam, portanto, pensamentos diferentes.

Do meu ponto de vista, estas objeções apoiam-se num holismo demasiado forte e incompatível quer com a (*RHOL*) de Davidson, quer com a sua tese sobre a autoridade de primeira pessoa[11]. É possível atribuir inteligivelmente *P* a *S* sem que *S* tenha, necessariamente, de acreditar que *Q*. Creio que isto não ameaça nem autoridade de *S* sobre o conteúdo de *P*, nem implica necessariamente que *S* e *Z* tenham pensamentos diferentes. É claro que *S* é, em parte, ignorante relativamente ao conteúdo do seu pensamento, mas esta ignorância não afeta nem a identidade da crença, nem a autoridade sobre o seu conteúdo. Penso que este é o caso pois, tal como Davidson (1997a, p. 124; 1982, p. 98) deixa assente em várias passagens, não há um conjunto fixo e determinado de crenças que um sujeito deva possuir de forma a se poder afirmar que acredita que *P*. Portanto, *S* precisa somente de acreditar que *P* implica, ou é implicada, por algumas crenças de *N*, sem necessariamente ter de acreditar que *P* implica, ou é implicada, por *Q*.

A fim de compreender o que levou Davidson a estipular a (*RHOL*) como critério para posse de crenças é conveniente regressar, por breves instantes, à (*RI*). Esta última, recorde-se, mostrou-nos que as atribuições *de dicto* não estão justificadas na ausência de linguagem, pois não nos permite fazer as "discriminações" que podem ser feitas quando as criaturas são capazes de enunciar respostas verbais. Por sua vez, a (*RHOL*) diz-nos que, ainda que aceitássemos as

[11] Davidson não apresenta nenhuma formulação exata da (*RHOL*), mas as suas afirmações segundo as quais a identidade de uma crença *P* requer a sua inserção num conjunto de crenças *N* no qual *P* se insere e mantém relações de coerência lógica parecem-me suportar esta interpretação. Veja-se Glock (1999) e Finkelstein (2007) para uma análise deste ponto.

atribuições *de re* expressas por (F4*) e (F5*) — ou seja, ainda que aceitássemos que Fido seria capaz de distinguir o objeto em causa ("árvore") sob uma hipotética, desconhecida, descrição —, teríamos de lhe atribuir o conceito "árvore" e, consequentemente, vários conceitos relacionados com este último. O problema que enfrentamos é que na ausência de respostas verbais por parte da criatura estas atribuições estão seriamente indeterminadas por quantos conjuntos de crenças sejamos capazes de imaginar. Se não conseguimos decidir qual conjunto de crenças aceitaria a criatura em questão, podemos igualmente duvidar se essa criatura terá de todo quaisquer crenças.

A (*RI*) e a (*RHOL*) encontram-se, assim, relacionadas. O argumento para a última pode ser reconstruído da forma que se segue:

(P1') *S* possui *P* somente se *S* possui um conjunto de crenças *N* com o qual *P* é coerente.

(P2') *S* possui um conjunto de crenças *N* somente se *S* é uma criatura linguística.

(P3') *S* não é uma criatura linguística.

(P3'') Portanto, *S* não possui um conjunto *N* de crenças.

(P4') Logo, *S* não possui *P*.

À semelhança da (*RI*), a (*RHOL*) constitui um argumento a favor de (P3). Mas constitui um argumento que pretende rejeitar (P1)? Vejamos:

> Provavelmente, estas considerações serão menos persuasivas para quem gosta de cães do que para quem não gosta, mas em qualquer caso *elas não constituem um argumento*. Na melhor das hipóteses, o que mostrámos, ou alegámos, é que a menos que haja comportamento que possa ser interpretado como linguístico, *os indícios não serão adequados para justificar as distinções finas que estamos acostumados a usar na atribuição de pensamentos*. (DAVIDSON, 1975, p. 164 – itálico acrescentado).

3.3 Conceito de crença

A restrição do conceito de crença assenta em duas suposições que Davidson (1982, p. 150) reconhece que podem ser colocadas em dúvida:

(i) A posse de crença requer a posse do conceito de crença.

(ii) A posse do conceito de crença requer a posse de linguagem.

Por conceito de crença, Davidson alude ao conceito de verdade objetiva, à capacidade de captar o contraste entre crenças verdadeiras e crenças falsas. A seu entender, se *S* possui de todo crenças, então *S* deverá ser capaz de captar a possibilidade de a sua crença ser falsa. Esta capacidade, segundo Davidson, surge somente com a linguagem.

A restrição do conceito de crença pode ser formulada do seguinte modo:

(*RCC*) *S* possui *P* somente se *S* reconhece que pode dar-se o caso de *P* [em *N*] ser falsa.

A (*RCC*) só pode ser compreendida em relação com as restrições anteriores. Para recordar, a (*RI*) e a (*RHOL*) mostram-nos, respetivamente, que as atribuições *de dicto* não estão justificadas na ausência de linguagem, e que as atribuições *de re*, ao exigirem a atribuição de uma ampla rede de crenças, estão drasticamente indeterminadas.

A fim de compreendermos a pertinência da (*RCC*) como condição necessária à posse de pensamento, consideremos os três cenários seguintes.

(*SC1*): Saio de casa pela manhã e fecho a porta. No momento em que fecho a porta, o telefone de casa começa a tocar e eu quero entrar para atender. Coloco a mão no bolso à procura das chaves da porta, mas não as encontro. Se não as encontrar, terei de arrombar a porta ou partir uma janela para entrar em casa. Em pânico, começo desesperadamente a procurar as chaves em todos os outros bolsos e na mochila que trago comigo até que, de repente, olho para o chão e as encontro. Aí percebi que em algum momento as teria deixado cair.

(*SC2*): Uma equipa de biólogos está a realizar testes experimentais em plantas de modo a estudar o seu comportamento. A planta em causa é um girassol. É sabido que o girassol manifesta um comportamento heliotrópico: um comportamento que se caracteriza pelo movimento de certas plantas em direção ao Sol. Numa destas experiências, os biólogos aproximam uma luz artificial forte ao girassol e este volta-se para ela, ignorando a luz solar.

(SC3): Recordemos os célebres estudos do comportamento dos macacos-vervet [*vervet monkeys*] que emitem diferentes sinais de alarme a fim de avisar os membros do seu grupo de situações de perigo, seja da aproximação de aves de rapina, de leopardos, de serpentes, ou de humanos (CHENEY & SEYFARTH, 1990; TOMASELLO, 1997; TOMASELLO, CALL & HARE, 2003). Suponhamos que, ao avistar leopardos nas proximidades, um macaco-vervet emite o sinal que é normalmente emitido na presença de águias. Como resposta a este sinal, os restantes membros do grupo manifestam um comportamento diferente daquele que usualmente se observa quando existem leopardos por perto: em vez de subirem às árvores, escondem-se nos arbustos mais próximos.

O que (*SC1*), (*SC2*) e (*SC3*) têm em comum é aquilo que descreveríamos como *erro*: em (*SC1*), a minha crença de que as chaves estavam no meu bolso era falsa; em (*SC2*), o girassol voltou-se para a luz artificial quando, em circunstâncias normais, o esperado seria que se voltasse para a luz solar; e em (*SC3*), o macaco-vervet emitiu um sinal de alarme de águia num contexto em que deveria ter emitido um sinal de alarme de leopardo. Contudo, (*SC1*) difere dos restantes neste sentido:

eu posso verbalizar, se me perguntarem, que a minha crença não correspondia ao estado real de coisas do mundo. E quanto a (*SC2*) e (*SC3*)? Será que podemos descrever o comportamento do girassol e do macaco-vervet como um comportamento de erro? Por que há de ser a capacidade de reconhecimento do erro essencial para a posse de pensamento? Para compreendê-lo, voltemos por uns breves momentos à (*RI*) mencionada anteriormente.

Uma resposta possível à (*RI*) poderia ser a seguinte: é suficiente, para justificar a atribuição de crença, que a criatura em questão *discrimine* objetos. Para isto, não é necessário que ela reconheça o objeto *sob certas descrições*, dado que a discriminação desse objeto pode ser levada a cabo em virtude das características perceptivas do mesmo. Por exemplo, Fido pode identificar o seu dono e distinguir-lo de outros indivíduos em virtude de determinadas características visuais ou olfativas. Assim, Fido pode acreditar que o seu dono chegou a casa sem ter de aceitar qualquer descrição do seu dono (por exemplo, que o seu dono é o Presidente dos EUA). Esta convicção, se defensável, parece levantar um desafio à tese davidsoniana de que a posse de conceitos é necessária à posse de crenças ou outras atitudes proposicionais.

Esta resposta, a meu ver, foca-se no lado errado do problema. A questão a discutir não é se uma criatura discrimina objetos em virtude das suas características perceptivas ou em virtude das suas descrições, mas antes qual a *natureza* dessa discriminação. É precisamente aqui que se justifica a entrada em cena da (*RCC*). Para Davidson, a atribuição de crença está justificada somente se a discriminação em causa for *normativa* ou *intencional*. A (*RCC*) introduz precisamente este factor normativo que permite ao atribuidor de crença saber quando a criatura está a agir ou a atuar *por uma razão* ou quando está apenas a manifestar uma atitude discriminatória e a atuar *como se fosse por uma razão*. Para a posse de pensamento "não é suficiente discriminar entre aspectos do mundo, comportar-se de diferentes maneiras em diferentes circunstâncias" (DAVIDSON, 1991, p. 209), visto que a performance discriminatória é comum até a plantas — e seguramente sentir-nos-íamos relutantes em atribuir crenças, por exemplos, a girassóis — e pode ser explicada como uma certa reação causal a um estímulo não requerendo, portanto, a atribuição de conteúdo proposicional. Para tornar possível efetuar esta distinção, Davidson introduz a necessidade de respostas verbais por parte da criatura.

O argumento da (*RCC*) pode ser reconstruído como se segue:

(P1#) Uma criatura *S* possui *P* se possui o conceito de crença.

(P2#) Para possuir o conceito de crença, *S* deve possuir linguagem.

(P3#) *S* não possui linguagem.

(P3#') Logo, *S* não possui o conceito de crença.

(P4#) Portanto, *S* não possui *P*.

Segundo a interpretação mais generalizada, é a (*RCC*) que permite fazer a distinção entre criaturas que são ou não pensantes, visto que nem a (*RI*) nem a (*RHOL*) são conclusivas. Em última análise, é também a (*RCC*) que mais pontua a

favor da *LF* e, consequentemente, a que mais apoia a negação de (P1). Eis onde eu penso que esta interpretação falha. A (*RCC*) a meu ver, permite-nos efetuar a distinção entre (P1) e (*UPAC*), i.e., proporciona-nos um critério que nos permite perceber quando as nossas atribuições de crença estão apenas *pragmaticamente justificadas* ou quando constituem uma *demonstração* da sua posse: estamos em posição de saber se uma criatura está simplesmente a reagir a um estímulo ou a manifestar um comportamento intencional se essa criatura for capaz de *indicar as razões* que motivaram esse comportamento. É claro que isto não significa que uma criatura só é capaz de agir mediante razões, ou só é capaz de comportamento intencional, se for capaz de as verbalizar. Esta não é, no entanto, a questão central. O objetivo aqui é saber quando estamos *justificados* a fazer atribuições de crença e quando constituem estas atribuições uma *demonstração* da presença das mesmas.

Posto isto, a meu entender nem a (*RCC*) nem as restrições anteriores pretendem rejeitar (P1), embora constituam argumentos fortes a favor de (P3). Uma das passagens de Davidson a favor da minha *LD* é, por exemplo, a seguinte:

> Poderíamos considerar esta linha, mas infelizmente parece não haver qualquer razão clara de por que temos de o fazer. Pretendíamos encontrar um argumento para mostrar que apenas criaturas com linguagem têm pensamentos. O que acaba de ser esboçado *não é um argumento, mas uma proposta, e uma proposta que não precisamos de aceitar*. (DAVIDSON, 1975, p. 167).

3.4 DISCRIMINAÇÃO E NORMATIVIDADE

Parece haver uma diferença óbvia entre (*SC2*) e (*SC3*) que nos faz sentir mais justificados a atribuir crenças em (*SC3*) do que em (*SC2*). Por exemplo, alguns estudos empíricos parecem suportar a conclusão de que os macacos-vervet manifestam uma conduta corretiva: os macacos-vervet mais experientes podem corrigir os macacos-vervet mais jovens que por vezes se enganam a emitir um dado sinal de alarme, sendo que esta correção permite-os melhorar a sua performance. É indiscutível que o comportamento dos macacos-vervet supera, quer em complexidade, quer em imprevisibilidade, o comportamento de girassóis. Mas será que este grau de complexidade acrescido nos torna mais justificados a acreditar que em (*SC3*) estamos perante criaturas pensantes ao passo que em (*SC2*) não?

Quando Davidson se refere às capacidades discriminatórias de criaturas não linguísticas, ele está a sugerir que estas podem ser descritas como resposta a um estímulo sem envolver a atribuição de conteúdo intencional: uma árvore, por exemplo, é capaz de discriminar solo seco de solo húmido, da mesma forma que Fido é capaz de discriminar o gato que persegue do seu dono. Isto pode parecer errado, pois há casos bem mais complexos de discriminação reportados em estudos empíricos que parecem levar-nos à conclusão de que pelo menos algumas criaturas não linguísticas são capazes de uma conduta autocorretiva. Voltemos por uns momentos ao cenário descrito em (*SC3*).

A observação do comportamento de macacos-vervet mais novos parece indicar que estes, por vezes, emitem sinais de alarme errados em determinada situação de perigo, sendo corrigidos por membros adultos. Esta conduta corretiva

permite que os membros mais jovens aprendam a melhorar a sua performance de modo a emitir o sinal de alarme adequado na presença de determinado animal. Ora, à primeira vista esta conduta pode ser considerada uma distinção pré-linguística, elementar, entre verdade e falsidade que emerge de percepções do meio ambiente e das respostas dessa criatura a determinados estímulos percetivos, não requerendo a posse de conceitos (BEISECKER, 2002, p. 118; MACINTYRE, 1999, p. 36; veja-se ainda ALLAN, 1999, pp. 36-37 e ALLAN & BEKOFF, 1996, p. 51). Mas estaremos justificados a atribuir-lhes crenças num sentido que vá mais além de (*UPAC*), i.e., num sentido que demonstre (P1)?

O problema que enfrentamos, para Davidson, prende-se com o facto de um dado comportamento poder parecer-nos errado *do nosso ponto de vista*. Talvez seja um comportamento errado do ponto de vista do membro da espécie, mas como poderemos saber se este é com efeito o caso? Se a criatura não é capaz de reconhecer o seu erro verbalmente, não poderemos sequer saber se o seu comportamento foi causado por um estímulo interno em vez de um estímulo externo. No caso descrito em (*SC3*), insistimos em explicar o comportamento de membros mais jovens da espécie fazendo referência à presença de águias nas redondezas, porque esta *nos parece* a explicação mais plausível. Mas como poderemos saber se este é o caso?

Na impossibilidade de responder a esta pergunta, as atribuições de crenças não são mais do que uma justificação pragmática do comportamento de criaturas. A posse de crença requer que a criatura possua a capacidade de se comportar não só de acordo com *uma regra*, mas que seja capaz de *seguir uma regra*: a discriminação entre *X*s e ¬*X*s, bem como o seu comportamento *X* e *Y*, devem estar motivados por *uma razão*. Segundo a minha interpretação, Davidson aceita que muito embora uma criatura seja capaz de agir por uma razão, de reconhecer o seu erro ou poder ter agido de outra forma, a menos que ela possua a capacidade de verbalizar as razões que estiveram na origem do seu comportamento, as suas discriminações não possuem a força normativa necessária para demonstrar (P1).[12] É por este motivo que Davidson afirma que

> [a] menos que queiramos atribui conceitos a borboletas e oliveiras, não devemos considerar a mera habilidade de discriminar entre vermelho e verde, ou húmido e seco, como tendo um conceito, nem sequer se tal comportamento seletivo for aprendido. (DAVIDSON, 1997b, p, 139).

[12] A meu ver, Davidson não está a negar que o comportamento de criaturas não linguísticas possa ser voluntário ou intencional. De igual modo, tal como o interpreto, não creio que ele acreditasse de facto que este comportamento pudesse ser explicado apelando única e exclusivamente a processos e vocabulário biológicos. Tanto crianças pré-linguísticas como algumas criaturas não linguísticas parecem capazes de realizar ações voluntárias, de se abster de executar uma determinada tarefa e de renunciar a determinados fins, ainda que temporariamente. De acordo com a *LD* que proponho, estas afirmações devem ser entendidas não como uma rejeição desta possibilidade, mas como uma falta de indícios que as garantam ou justificam. Onde Davidson falha, se assim se pode dizer, é em não distinguir criaturas pré-linguísticas de criaturas não linguísticas. No entanto, se entendermos que os seus argumentos que apresentei em §4.1, §4.2 e §4.3 põem em relevo a falta de indícios para demonstrar (P1), mas não negam a (*UPAC*), embora a importância desta distinção não desapareça, em certa medida, desvanece.

Este ponto de vista pode ser pouco convincente, mas penso que nos ajuda a compreender por que motivo Davidson (1982, p. 95) afirma que ""nem uma criança de uma semana, nem um caracol, é uma criatura pensante". Ainda que o comportamento dos macacos-vervet seja compatível com a explicação de que estes manifestam uma conduta corretiva, uma das dificuldades em justificar as atribuições de crença como demonstrando (P1) reside na dificuldade em garantir que o seu comportamento se trata efetivamente de um comportamento intencional, e não de uma mera resposta discriminatória a um dado estímulo. Ainda que os macacos-vervet mais jovens pareçam ser corrigidos por elementos mais velhos e possam aprender a melhorar a sua performance na presença de determinadas situações de perigo, permanece aberta a questão de saber se estes de facto *reconhecem* e *corrigem* os seus *erros* ou se, simplesmente, adaptam o seu comportamento. O problema com que nos deparamos é que não conseguimos saber se o seu comportamento é causal ou intencional, pois não fica claro se adquiriram apenas mais capacidades cognitivas (i.e., se ganharam mais informação ou conhecimento) ou se adquiriram capacidades para avaliar e exercer juízos.

Por este motivo, apesar de estarmos justificados a atribuir crenças aos macacos-vervet, estas justificação não superaria a justificação pragmática aludida em (*UPAC*). O que isto significa, nas palavras de Davidson é o seguinte:

> Para explicar o comportamento dos macacos não precisamos de lhes atribuir intenções ou crenças (*não estou a argumentar que eles não têm intenções ou crenças*). E assim nada no seu comportamento como descrito tem de contar como cometer um erro. Evidentemente, pode acontecer que um macaco solte o "grito de perigo" quando o perigo não está presente, e os seus companheiros possam reagir ao perigo. Do nosso ponto de vista, isto pode parecer um erro. Mas a menos que os macacos acreditem que há perigo quando não há, nenhum erro foi cometido; eles simplesmente responderam a um estímulo que frequentemente, mas não sempre, acompanha o perigo. (DAVIDSON, 1987, p. 116 – itálico acrescentado).

Quando, em várias ocasiões, Davidson (por exemplo, 1982, p. 105; 1982b, p. 100) diz que a linguagem é um traço social, ele não está a negar que criaturas não linguísticas possuam pensamentos. Ou seja, Davidson não está a rejeitar (P1). Os seus argumentos são compatíveis com a *LD* segundo a qual ao não possuirmos uma base sólida para pelo menos atenuar (P3), não estamos em posição demonstrar (P1) — e não estar em posição de demonstrar (P1) não é o mesmo que rejeitar (P1)! —. O que estudos empíricos nos mostram é que as observações do comportamento de uma determinada criatura não linguística não são suficientes, *de um ponto de vista conceptual*, para demonstrar a posse de crenças (veja-se também DRECKMAN, 1999). Contudo, nada disto significa que estas criaturas não as possuam. Na sequência da citação anterior, Davidson acrescenta algumas palavras que me parecem suportar a minha leitura:

> Considero óbvio que o comportamento linguístico é intencional e, por isso, requer crença. O conceito de erro pode ser aplicado somente quando intenção e crença estão presentes. Na nossa

> história, as reações do macaco não são intencionais, e portanto, obviamente, não podem ser um modelo para o comportamento linguístico. Mas este é apenas uma parte da questão em causa. A questão em causa pode ser melhor colocada desta maneira: *esteja ou não a intenção presente, não há elementos suficientes para garantir que esteja*. (DAVIDSON, 1987, p. 116 – itálico acrescentado).

4 Triangulações, ação e indeterminação

Vimos até aqui que os argumentos de Davidson tipicamente tomados como um suporte para a *LF* baseiam-se no facto de na ausência de respostas verbais por parte da criatura as nossas atribuições de crença *de dicto* não são conclusivas, pois não podemos fazer as "discriminações finas" que tipicamente fazemos quando a criatura é capaz de enunciar respostas verbais, e as nossas atribuições de crença *de re* estão drasticamente indeterminadas, visto que podemos atribuir-lhe quantos sistemas de crenças sejamos capazes de imaginar, o que nos deixa sem bases para saber se a criatura está a manifestar um comportamento intencional ou causal, i.e., se está a agir por uma razão ou a responder a um estímulo cuja localização não temos como determinar.

Uma forma de examinar a relação pensamento-linguagem consiste em analisar a teoria que subjaz ao tipo de explicações que usamos para esclarecer, precisamente, essa relação. Parte dessa teoria "lida com a explicação teleológica da ação" (DAVIDSON, 1975, pp. 158-159). Uma das características da explicação da ação não partilhada por outros tipos de explicações de fenómenos físicos que observamos — por exemplo, ver uma pedra cair — consiste no facto de estas apelarem ao conceito de razão exibindo, assim, "a racionalidade da ação à luz do conteúdo da crença e do objeto do desejo" (DAVIDSON, 1975, p. 159).

A dificuldade que surge em atribuir crenças a criaturas apenas através da observação do comportamento não verbal e tomar essa atribuição como uma demonstração da posse das mesmas não reside no facto de ser difícil "descobrir essas coisas na ausência de linguagem, mas no facto de não termos qualquer boa ideia de como começar a *autenticar* a existência dessas atitudes quando a comunicação não é possível" (DAVIDSON, 1974, p. 114 – itálico acrescentado). Se assumirmos que existe uma distinção entre ações intencionais e o comportamento adaptativo a circunstâncias do ambiente, torna-se necessário estipular um critério que nos permita diferenciar ambos. Davidson recorre à linguagem. Podemos discordar deste critério, mas esta discórdia não constitui por si uma razão a favor da *LF*. Afinal, por que motivo Fido só pode possuir a crença de que o gato trepou àquela árvore se for capaz de verbalizar essa crença? O problema é que ao explicar o comportamento de Fido através da atribuição de crenças, o que estamos a fazer é apenas uma suposição (bastante plausível) acerca da finalidade desse comportamento. É por este motivo que Davidson defende que "a atribuição de desejos e crenças (e outros pensamentos) deve andar de mãos dadas com a interpretação da fala" (1975, p. 163). Se nos perguntarmos por que não eliminar

simplesmente a necessidade de resposta verbais por parte da criatura, Davidson responde:

> É relativamente simples eliminar a necessidade de respostas verbais por parte do sujeito: pode considerar-se que ele expressou uma preferência ao realizar uma ação, movendo-se diretamente de modo a alcançar o seu objetivo, em vez de dizer o que quer. *Mas isto não pode resolver a questão acerca do que escolheu.* Um homem que pega numa maçã em vez de numa pera quando se lhe oferecem ambas, pode estar a expressar uma preferência pelo que está à sua esquerda em vez do que está à sua direita, pelo que é vermelho em vez de amarelo, pelo que viu em primeiro lugar, ou julga que é mais caro. Testes repetidos podem fazer algumas leituras das suas ações mais plausíveis do que outras, *mas o problema acerca de como determinar quando ele julga que dois objetos de escolha são idênticos permanecerá.* Os testes que envolvem certos acontecimentos — escolhas entre apostas — são ainda mais difíceis de apresentar sem usar palavras. O psicólogo, cético quanto à sua habilidade de estar certo sobre como um sujeito está a interpretar as suas instruções, tem de adicionar uma teoria de interpretação verbal à teoria a ser testada. Se pensarmos que todas as escolhas revelam uma preferência que uma frase, em vez de outra, é verdadeira, a teoria resultante total deve fornecer uma interpretação das frases, e ao mesmo tempo atribuir crenças e desejos, ambos concebidos como relacionando o agente a frases ou enunciações. Esta teoria composta explicaria todo o comportamento, verbal ou não. (DAVIDSON, 1975, pp. 162-163 – itálico acrescentado).

Com isto, Davidson não está a dizer que uma criatura só age por uma razão se for capaz de verbalizar essa razão; o que ele diz é que na ausência de linguagem a existência ou a escolha dessa razão não tem como ser determinada e, por conseguinte, não temos como escolher entre variadas hipóteses explicativas igualmente válidas. Isto vale tanto para criaturas não linguísticas, pré-linguísticas, como para criaturas linguísticas que não verbalizam as suas respostas.

Davidson explica a diferença aqui em causa, entre criaturas linguísticas e não linguísticas, mediante a metáfora da triangulação. Trata-se de uma metáfora que, muito mais do que explicar como as palavras adquirem o seu significado, consiste numa proposta explicativa de como o pensamento e a linguagem surgiram. Davidson (1997a) fala-nos em dois tipos de triangulação: uma triangulação pré-linguística ou pré-cognitiva, e uma triangulação linguística e cognitiva, sendo que na primeira o pensamento e a linguagem ainda não estão presentes, ao passo que na segunda as criaturas são já linguísticas e pensantes. Descrever a passagem entre o primeiro e o segundo tipo de triangulação equivaleria a explicar a emergência do pensamento e da linguagem. No entanto, segundo Davidson (1997a, p. 127), esta passagem não pode ser descrita visto que carecemos de um vocabulário satisfatório para explicar estágios intermédios.

Não há nenhum segredo, de acordo com Davidson, para interpretar e dar sentido ao comportamento linguístico de criaturas: é preciso observar o seu comportamento linguístico e não linguístico, olhar para o mundo e correlacionar

esse comportamento com acontecimentos que se encontram lá fora, mais além da sua pele. Não há qualquer mistério para interpretar o comportamento de falantes. Ao observar o comportamento de um homem que puxa, em direções opostas, dois extremos de uma corda podemos interpretar o seu comportamento de várias formas.[13] Podemos, entre várias explicações possíveis, considerar que está a lutar consigo mesmo e pretende saber qual dos lados sairá vitorioso ou que está a testar a elasticidade da corda. Se ele é um falante competente — i.e., um falante capaz de interagir linguisticamente com outros falantes através da interpretação e enunciação de frases —, poderemos facilmente decidir entre estas duas opções perguntando-lhe o que pretende fazer. Ao ouvir a sua resposta de que pretende cortar a corda em dois, percebemos que as nossas hipóteses explicativas prévias, apesar de plausíveis, estavam erradas. Mas o que fazer em cenários em que não podemos ter acesso às respostas que explicariam o comportamento que queremos interpretar?

A resposta de Davidson é que, neste caso, as hipóteses explicativas que temos para esse comportamento seriam demasiado abundantes, deixando o mesmo drasticamente indeterminado.[14] No fundo, não teríamos como decidir se a criatura em questão está a responder intencionalmente a um estímulo externo, situado no mundo intersubjetivamente comum, ou se está a manifestar uma mera disposição causal provocada por um estímulo interno, localizado no seu corpo e inacessível de um ponto de vista de terceira pessoa. Como consequência, não haveria lugar para a normatividade requerida nas explicações e que nos permite explicar o conceito de verdade objetiva, isto é, os possíveis cenários de erro mencionados anteriormente. Na ausência de respostas verbais por parte da criatura, qualquer comportamento será correto se nos parecer, *do nosso ponto de vista e não do ponto de vista da criatura*, correto. A entender de Davidson, isto deixa de contar como uma hipótese explicativa genuína, passando a contar como uma mera descrição de um comportamento. Do seu ponto de vista, só a presença de objetos observados conjuntamente permite, mediante interações linguísticas, fixar e identificar corretamente o conteúdo de um pensamento.[15]

[13] Retomo este exemplo de Davidson (1975).

[14] A seguinte objeção, extremamente pertinente, foi levantada por um parecerista. Poder-se-ia dizer que este ponto é redundante, visto que a indeterminação está também presente na interpretação do comportamento linguístico. A ameaça aparente para o meu argumento é a seguinte: fundamentar um ceticismo quanto à posse de crenças na indeterminação da explicação do comportamento de criaturas não linguísticas levar-me-á forçosamente a aceitar um ceticismo quanto à posse de crenças em criaturas linguísticas. No fundo, a indeterminação é transversal à explicação de todo o tipo de comportamentos, verbais e não verbais, e a conclusão do cenário da intepretação radical. Este é, reconhecidamente, um ponto sensível deste artigo. Tudo o que eu posso dizer aqui é, talvez, que no caso da interpretação linguística o ceticismo recai sobre um certo número de *manuais de interpretação*, igualmente válidos, para o mesmo comportamento, ao passo que na interpretação do comportamento não verbais o ceticismo recai sobre *a posse ou não*, literal, de crenças por parte das criaturas em questão. Para apoiar esta resposta terei de dizer que a indeterminação não é a última palavra sobre o ceticismo ou não ceticismo quanto à posse de crenças, embora eu acredite — e muitas passagens de Davidson o suportem, a meu ver — que ela desempenhe um papel crucial no que procuro argumentar. Outro critério igualmente inabdicável, e que deverá ser mencionada a par com a drástica indeterminação explicativa na ausência de linguagem, é a posse do conceito de crença mencionado em §4.3.

[15] Para uma explicação mais pormenorizada deste ponto veja-se §5.2.1 do meu livro (COUTO, 2018).

A dificuldade que enfrentamos quando procuramos explicar o comportamento de criaturas não linguísticas surge igualmente quando procuramos explicar o comportamento de humanos adultos *apenas* através da observação do seu comportamento não linguístico. Segundo Davidson, é por este motivo que não podemos ter uma teoria da ação sem uma teoria da interpretação, pois na ausência de linguagem "Mesmo uma amostra generosa de ações ameaça deixar em aberto *um número inaceitavelmente grande de explicações alternativas*" (DAVIDSON, 1975, p. 160-161 – itálico acrescentado). Em última análise é por este motivo que:

> Temos a ideia da crença apenas do papel desta na interpretação da linguagem, pois enquanto atitude privada ela não é inteligível a não ser como um ajuste à norma pública fornecida pela linguagem. Segue-se que uma criatura tem de ser membro de uma comunidade linguística para ter o conceito de crença. E dada a dependência de outras atitudes em relação à crença, podemos dizer de maneira mais geral que somente uma criatura que pode interpretar a fala pode ter o conceito de um pensamento. (DAVIDSON, 1975, p. 170).[16]

A falta de garantias que acompanha as nossas atribuições de crença *de re* e *de dicto* faz com que "as nossas atribuições e consequentes explicações de ações estarão seriamente indeterminadas no sentido em que muitos sistemas de atribuição alternativos, muitas explicações alternativas, estarão igualmente justificadas pelos dados disponíveis" (DAVIDSON, 1975, p. 164). Se insistirmos em atribuir crenças na ausência de respostas verbais por parte da criatura, as nossas atribuições e consequentes explicações ver-se-ão seriamente indeterminadas por quantos sistemas alternativos de atribuição de crenças igualmente válidos possamos imaginar. A grande dificuldade em atribuir crenças na ausência de linguagem não é que tais atribuições não possam ser feitas ou que sejam meramente instrumentais. Em suma, a dificuldade em justificar as atribuições de crença na ausência de linguagem não equivale, nem pretende, rejeitar (P1):

> Atitudes proposicionais podem ser descobertas por um observador que não observa nada além de comportamento sem que as atitudes sejam de alguma forma redutíveis ao comportamento. Existem laços conceituais entre as atitudes e o comportamento que, com suficiente informação sobre o comportamento real e potencial, são suficientes para permitir fazer inferências de atitudes. (DAVIDSON, 1982, p. 100).

Conclusão

Neste artigo, sugeri uma interpretação não ortodoxa da abordagem de Davidson às mentes animais que não o toma como um defensor estrito do lingualismo. Segundo esta interpretação — a que chamei *LD* — as polémicas afirmações de Davidson que à primeira vista suportam a *LF* — por exemplo: "uma criatura não pode ter pensamentos a menos que tenha uma linguagem", "para ser uma criatura racional pensante, a criatura deve ser capaz de expressar muitos

[16] Em última análise, é também por isto que Davidson (1980) afirma que a teoria da interpretação deve avançar em direção a uma teoria unificada do pensamento, do significado, e da ação.

pensamentos e, acima de tudo, ser capaz de interpretar a fala e os pensamentos de outros" (1982, p. 100), "uma criatura não pode ter pensamentos, a menos que seja intérprete da fala de outra" (1975, p. 157) — constituem argumentos a favor de (P3), mas não negam (P1). O que proponho com a *LD* é que estas afirmações devem ser entendidas não como uma negação de (P1), mas como ceticismo quanto a (P1). Este ceticismo tem as suas raízes na indeterminação explicativa do comportamento de criaturas não linguísticas e não na convicção ou crença de que estas não possuem, ou não podem possuir, crenças.

É certo que nada disse sobre a premissa verificacionista que mencionei em §2. Muitos leitores poderão alegar que a *LD* que proponho não livra Davison da denúncia de que a sua postura perante a relação entre linguagem e pensamento implica um certo verificacionismo. Talvez isto seja verdade. No entanto, este será um verificacionismo fraco que em lugar de defender que "devemos *acreditar* apenas no que podemos verificar", assenta na ideia de que "podemos *demonstrar* somente o que podemos verificar". Esta é uma interpretação que, pelo menos, me parece consistente com os argumentos de Davidson. Se (P3) é verdadeira, então não temos fundamentos para demonstrar a verdade, ou falsidade, de (P1). É desta forma que devemos interpretar a conclusão de Davidson de que "racionalidade é um traço social" e que "somente quem se comunica a tem" (1982, p. 105).

Davidson reconhece em várias passagens que talvez pudesse ser possível avançar uma teoria da crença que não fizesse referência à linguagem, apesar de a crença em si mesma só ser inteligível por referência à linguagem. A meu ver, estas passagens foram negligenciadas pela grande parte dos intérpretes de Davidson. Contra a ideia generalizada que o associa à *LF*, o que Davidson defende é somente que "somente homens e mulheres têm linguagem, ou algo bastante parecido a uma linguagem, para justificar a atribuição de pensamentos preposicionais" (1982, pp. 96, fn.1). Isto parece-me algo muito diferente de rejeitar *tout court* esta possibilidade. A *LD* que sugiro pretende assim dar ênfase às seguintes palavras de Davidson:

> A suposição frequente é que um ou outro, linguagem ou pensamento, é, por comparação, fácil de compreender, e portanto o mais obscuro (qualquer que seja) pode ser iluminado analisando-o ou implicando-o em termos do outro. Esta suposição é, creio, falsa: *nem a linguagem, nem o pensamento, podem ser inteiramente explicados em termos do outro, e nenhum tem prioridade conceptual.* Os dois estão, com efeito, ligados, no sentido de que cada um requer o outro de modo a ser compreendido. (DAVIDSON, 1975, p. 156 – itálico acrescentado).

Referências

ALLEN, Colin. Animal concepts revised: the use of self-monitoring as an empirical approach. *Erkenntnis,* v. 51, pp. 33-40, 1999.

ALLEN, Colin; BEKOFF, Marc. Intentionality, social play and definition. In: BEKOFF, M.; JAMIESON, D. (Eds.) *Readings in animal psychology.* Cambridge, MA: MIT Press, 1996. pp. 229-239.

_____. *Species of mind*: the philosophy and biology of cognitive ethology. Cambridge, MA: MIT Press, 1997.

BARTH, Christian. *Objectivity and the language-dependence of thought*: a transcendental defense of universal lingualism. New York; London: Routledge, 2011.

BEISECKER, David. Some more thoughts about thought and talk: Davidson and fellows on animal beliefs. *Philosophy,* v. 77, pp. 115-124, 2002.

CHENEY, Dorothy.; SEYFARTH, Robert. *How monkeys see the world*: inside the mind of another species. Chicago: University of Chicago Press, 1990.

COUTO, Diana. *Donald Davidson: subjetivo-objetivo. O retorno ao cogito*. Col. MLAG discussion papers, v. 10. Porto: Faculdade de Letras da Universidade do Porto, 2018.

DAVIDSON, Donald. Radical interpretation. In: DAVIDSON, Donald. *Inquiries into truth and interpretation.* Oxford: Oxford University Press, 1973. pp. 125-139.

_____. Belief and the basis of meaning. In: DAVIDSON, Donald. *Inquiries into truth and interpretation.* Oxford: Oxford University Press, 1974. pp. 141-154.

_____. Thought & Talk. In: DAVIDSON, Donald. *Inquiries into truth and interpretation.* Oxford: Oxford University Press, 1975. pp. 155-170.

_____. A unified theory of thought, meaning, and action. In: DAVIDSON, Donald. *Problems of rationality*. Oxford: Oxford University Press, 1980. pp. 151-166.

_____. Rational animals. In: DAVIDSON, Donald. *Subjective, intersubjective, objective.* Oxford: Oxford University Press, 1982. pp. 95-105.

_____. Problems in the explanation of action. In: DAVIDSON, Donald. *Problems of rationality.* Oxford: Oxford University Press, 1987. pp. 101-116.

_____. The myth of the subjective. In: DAVIDSON, Donald. *Subjective, intersubjective, objective.* Oxford: Oxford University Press, 1988. pp. 39-52.

_____. Three varieties of knowledge. In: DAVIDSON, Donald. *Subjective, intersubjective, objective.* Oxford: Oxford University Press, 1991. pp. 205-220.

_____. The emergence of thought. In: DAVIDSON, Donald. *Subjective, intersubjective, objective.* Oxford: Oxford University Press, 1997a. pp. 123-134.

_____. Seeing through language. In: DAVIDSON, Donald. *Truth, language, and history.* Oxford: Oxford University Press, 1997b. pp. 127-141.

_____. Indeterminism and antirealism. In: DAVIDSON, Donald. *Subjective, intersubjective, objective.* Oxford: Oxford University Press, 1997c. pp. 69-84.

DAVIDSON, Donald. Interpretation: hard in theory, easy in practice. In: CARO, Mario (ed.) *Interpretations and causes*: new perspectives on Donald Davidson's philosophy. Dordrecht: Kluwer Academic Publishers, 1999. pp. 31-44.

_____. *Truth and predication.* Cambridge, MA: Harvard University Press, 2005.

DENNETT, Daniel. Intentional systems in cognitive ethology: The 'Panglossian Paradigm' defended. *Behavioral and Brain Sciences,* v. 6, pp. 343-390, 1983.

DRECKMANN, Frank. Animal beliefs and their contents. *Erkenntnis,* v. 51, pp. 93-111, 1999.

FELLOWS, Roger. Animal beliefs. *Philosophy - Cambridge University Press on behalf of Royal Institute of Philosophy,* v. 75, n. 294, pp. 587-598, 2000.

FINKELSTEIN, David. Holism and animal minds. In: CRARY, Alice. (ed.), *Wittgenstein and the moral life*: essays in honor of Cora Diamond. Cambridge, MA: MIT Press, 2007. pp. 251-278.

_____. Animal, thoughts and concepts. *Synthese,* v. 123, n. 1, pp. 35-64, 2000.

GLOCK, Hans-Johann. Animal minds: conceptual problems. *Evolution and Cognition*, v. 5, n. 2, p.174-188, 1999.

_____. *Quine and Davidson on language, thought and reality.* Cambridge: Cambridge University Press, 2003.

GLÜER, Kathrin. *Donald Davidson*: a short introduction. Oxford: Oxford University Press, 2011.

JAMIESON, Dale. What do animals think? In: LURZ, Robert W. (ed.) *The philosophy of animal minds.* Cambridge: Cambridge University Press, 2009. pp. 15-51.

LOWE, E. J. Personal experience and belief: the significance of external symbolic storage for the emergence of modern human cognition. In: SCARREE, Chris; RENFREW, Colin (eds.) *Cognition and culture*: the archhaeology of symbolic storage. Cambridge: McDonald Institute for Archaeological Research, 1998. pp. 89-96.

_____. *An introduction to the philoosphy of mind.* Cambridge: Cambridge University Press, 2004.

MACINTYRE, Alasdair. Can animals without language have beliefs? In: MACINTYRE, Alasdair. *Dependent rational animals*: why human beings need the virtues. Chicago: Open Court, 1999. pp. 29-41.

MALCOLM, Norman. Thoughtless brutes. *Proceedings and Addresses of the American Philosophical Society,* v. 46, pp. 5-20, 1972-3.

MIGUENS, Sofia. Conceito de crença, triangulações e atenção conjunta: Donald Davidson e a ciência cognitiva II. In: MIGUENS, S. *Será que a minha mente está dentro da minha cabeça?* da ciência cognitiva à filosofia. Porto: Campo das Letras, 2008 [2006], pp. 131-143.

MIGUENS, Sofia. Por que não pode haver uma ciência da racionalidade: Davidson e a ciência cognitiva I. In: MIGUENS, Sofia. *Será que a minha mente está dentro da minha cabeça – Da ciência cognitiva à filosofia*. Porto: Campo das Letras, 2008, pp. 121-129.

_____; PINTO, João Alberto. Seeing what a 'science of rationality' founders on (with a little help from D. Davidson). *Poznan Studies in the Philosophy of the Sciences and the Humanities*, v. 111, pp. 71-92, 2018.

PINKER, Steven. *The language instinct.* Middlesex: Penguin, 1994.

SEYFARTH, Robert; CHENEY, Dorothy. Inside the mind of a monkey. In: ALLEN, Colin; JAMISON, Dale. (eds.) *Readings in animal cognition.* Cambridge, MA: MIT Press, 1996. pp. 337-343.

_____; MARLER, P. Monkeys responses to three different alarm calls: evidence of predator classification and semantic communication. *Science,* v. 210, n. 4471, pp. 801-803, 1980.

STICH, Stephen. Do animals have beliefs? *Australasian Journal of Philosophy,* v. 57, n. 1, pp. 15-28, 1979.

TOMASELLO, Michael. *Primate cognition.* Oxford: Oxford University Press, 1997.

_____; CALL, Josep; HARE, Brian. Chimpanzees understand psychological states – the question is which ones and to what extent. *Trends in Cognitive Science,* v. 7, pp. 153-156, 2003.

WILLIAMS, Bernard. Deciding to Believe. In: WILLIAMS, Bernard. *Problems of the self*: philosophical papers, 1956-1972. Cambridge: Cambridge University Press, 1973. pp. 136-152.

WILLIAMSON, Timothy. Philosophical 'intuitions' and scepticism about judgement. *Dialectica,* v. 58, pp. 109-153, 2004.

Recebido em: 05-03-2019

Aceito para publicação em: 25-06-19

SO LANGUAGE. VERY PRESCRIBE. WOW

TANTO LINGUAGEM. BASTANTE PRESCREVER. UAU

SHANE NICHOLAS GLACKIN[1]
University of Exeter – UK
S.N.Glackin@exeter.ac.uk

ABSTRACT: The philosophical dispute about linguistic normativity is one battlefield in a larger war over the nature of language as an object of scientific study. For those influenced by Wittgenstein, language involves following – or failing to follow – public, prescriptive rules; for Chomsky and his followers, language is a property of individual minds and brains, and the grammatical judgements of any mature individual speaker – her competence – cannot be, in any linguistic sense, "wrong". As I argue here, the recent "doge meme" internet fad provides surprising evidence for the prescriptivist view. Normative attitudes towards linguistic practices are a ubiquitous feature of those practices, and there is no principled basis on which to regard them as non-linguistic.

KEYWORDS: Chomsky. Doge. I-Language. Normativity. Prescriptivism.

RESUMO: *A disputa filosófica sobre a normatividade linguística é um campo de batalha em uma guerra maior sobre a natureza da linguagem como objeto de estudo científico. Para aqueles influenciados por Wittgenstein, a linguagem envolve seguir - ou deixar de seguir - regras públicas, prescritivas; para Chomsky e seus seguidores, a linguagem é uma propriedade de mentes e cérebros individuais, e os julgamentos gramaticais de qualquer falante individual maduro - sua competência - não podem ser, em qualquer sentido linguístico, "errados". Como argumento aqui, a recente moda da Internet "meme doge" fornece evidências surpreendentes para a visão prescritivista. Atitudes normativas em relação às práticas linguísticas são uma característica onipresente dessas práticas, e não há nenhuma base bem fundamentada sobre a qual considerá-las como não linguísticas.*

PALAVRAS-CHAVE: *Chomsky. Doge. Língua-I. Normatividade. Prescritivismo.*

The familiar popular dispute between descriptivists and prescriptivists about grammar is – let us not mince words here – a profoundly tiresome and pointless one (WALLACE, 2001), which the explosion of opportunities for pedantry afforded by the internet has done everything to amplify, and nothing to revivify; I have no intention of pursuing it here. But the dispute has roots, or at any rate analogues, in some deep issues in theoretical linguistics and the philosophy of language; and as I shall argue, interesting light is shed on *that* dispute by a most unlikely online phenomenon.

I begin by summarising the main threads of the philosophical debate about normativity in grammar, and the apparent stalemate they have recently issued in. I then try to make whatever sense can be made of the wilfully absurd "Doge" meme,

[1] PhD in Philosophy (University of Leeds, 2008). Lecturer at the Department of Sociology and Philosophy - University of Exeter, UK.

before showing how that phenomenon, surprisingly, provides significant evidence for the prescriptivist case.

1

The philosophical dispute about linguistic normativity is one battlefield in a larger war over the nature of language as an object of scientific study. According to Noam Chomsky and his followers, the popular view of language as a public, collective, or abstract entity has no place in a scientific worldview; language is a feature of the individual brains of individual speakers. And one consequence of this view is that mature speakers cannot be 'wrong' in their linguistic judgements or practices, by any scientific standard; linguistic should aim to describe the *I-language* – the relevant features of individuals' minds and brains – and the behaviour that results from them, and eschew any attempt to evaluate that behaviour as proper or improper.

The radicalism of the I-language or "mentalist" position (ISAC & REISS, 2008, p. 12) is often unappreciated. "There is simply no way of making sense of [a notion of 'common, public language']," writes Chomsky, "… or of any of the work in theory of meaning and philosophy of language that relies on such notions" (1995, p. 48-49). And lest we thought he was making a narrower claim than he appears to be in this statement, he immediately assures us that it "is intended to cut rather a large swath" (*ibid.*). Variations on this theme recur throughout Chomsky's writing. Whereas the I-language is "a real object of the real world" (CHOMSKY, 1993, p. 39), and "as real as chemical compounds" (CHOMSKY, 1988, p. 679), public language ('E-Language') is "arbitrary" and "artifactual" (CHOMSKY, 1986, p. 26); despite philosophers' "constant reliance on some notion of 'community language' or 'abstract language', there is virtually no attempt to explain what it might be" (CHOMSKY, 1993, p. 39). And public language is "useless for any form of theoretical explanation" (CHOMSKY, 1995, p. 48), playing no "role in an eventual science of language" (CHOMSKY, 1986, p. 16); we "gain no insight into what [language-learners] are doing by supposing that there is a fixed entity that they are approaching, even if some sense can be made of this mysterious notion" (CHOMSKY, 1992a, p. 17). Any distinctions we might wish to draw between different "public languages" are therefore matters of class, politics, or race, rather than linguistics; "(p)eople who live near the Dutch border", he writes, "can communicate quite well with people living on the German side, but they speak different languages in accordance with the sense of the term [Michael] Dummett argues is 'fundamental'" (CHOMSKY, 1992b, p. 101).

On Chomsky's view, then, there are no such entities, abstract or concrete, as "public languages"; at most, there are mereological sums of more-or-less overlapping particular I-languages, embodied in the brains of particular individuals. But these sums play no explanatory or theoretical role in linguistic science, and there is no scientific basis for drawing their boundaries in one place rather than another.

Now, if language is not a property of communities, as Ludwig Wittgenstein recognised, then there cannot be any question of any individual speaking "correctly" or "incorrectly". If there is no external, community standard by which to judge my speech, "whatever is going to seem right to me is right. And that only means that here we can't talk about 'right'" (WITTGENSTEIN, 1953, §258). For Wittgenstein it followed that, language being normative, its rules must be matters of public convention. But one man's modus ponens is another's modus tollens; for Chomsky, it followed that language, being a property of individual minds and brains, could not have rules in Wittgenstein's sense at all (*cf.* CHOMSKY, 2013, p. 183).

It is to this issue, indeed, that we can perhaps trace Chomsky's introduction of the terms *I-Language* and *E-Language*. Chomsky had previously distinguished between *competence* and *performance*; the linguistic knowledge possessed by a speaker and the concrete linguistic phenomena they produced as a result. But Saul Kripke's exegesis of Wittgenstein's rule-following considerations directly challenged this terminology; competence, he observed, is "not a dispositional notion. It is normative, not descriptive… [it] is dependent on our understanding of the idea of 'following a rule'" (1982, p. 31, fn. 22). "Modern transformational linguistics", as a result, "inasmuch as it explains all my specific utterances by my 'grasp' of syntactic and semantic rules generating infinitely many sentences with their interpretation, seems to give an explanation of the type Wittgenstein would not permit" (*ibid.*, p. 97, fn. 77).

So much the worse, then, in Chomsky's eyes, for Wittgenstein. In his first major work following the publication of Kripke's lectures (which had circulated for several years previously) Chomsky not only responds to "Wittgenstein",[2] but also replaces the older competence/performance distinction with the new I-Language/E-Language one. It is in his rejoinder to the Wittgensteinian criticism that Chomsky first explicitly denies that there are "rules" of language in any familiar sense (CHOMSKY, 1986, p. 688 *ff.*). Language consists not of normative rules, but of purely descriptive principles and parameters for neural and mental organisation, and these are to be understood simply as natural, biological features typical to our species (*ibid.*; *cf.* CHOMSKY, 1995, p. 17).[3] The underlying dichotomy presumed here between the normative and the natural has been forcefully challenged (MILLIKAN, 1995; MILLIKAN, 2003; MILLIKAN, 2005); nevertheless, for Chomsky, normativity and prescription "plainly has nothing to do with an eventual science

[2] It is to the version of Wittgenstein presented by Kripke – often referred to as "Kripkenstein" or "Kripke's Wittgenstein" – that he in fact responds; the consensus view seems to be that this does *not* represent the real Wittgenstein's position, and Kripke is careful not to assert that it does (STEINER, 2011, p. 170, fn. 20). McNally & McNally (2012) provides a useful overview of Chomsky's response.

[3] One notable failure to appreciate the radicalism of Chomsky's turn here can be found in Devitt (2008), whose index contains a single combined entry for "rules (or principles)", and generally treats this as a merely terminological shift, attributing a single position on the status of linguistic rules and their mental representation to Chomsky on the basis of both pre- and post-1986 writings (*e.g.* pp. 04; 69; 174-175). But Chomsky has tended to obscure the discontinuity of his views here, and I myself have previously overlooked the novelty of the I-Language position (GLACKIN, 2011, p. 203; p. 210).

of language, but involves other notions having to do with authority, class structure, and the like" (CHOMSKY, 1988, p. 675; c*f.* ISAC & REIS, 2008, ch. 12).

According to Chomsky, then, language is not the sort of thing mature individual users can "get wrong"; at most, we can descriptively state that they fail to make themselves understood (CHOMSKY, 2000, p. 07). However, this view not only runs counter to those of the great number of philosophers influenced by Wittgenstein; it is also at odds with the phenomenology of our language-use.

It is uncontroversial that we do in fact experience the "rules" of our grammar, whether or not we can formulate them explicitly, as having some sort of normative pull. "When we hear ['The child seems sleeping']", write the authors of a recent pro-Chomsky textbook, "we automatically interpret it as meaning basically the same thing as what *The child seems to be sleeping* means. And yet, it intuitively feels like there is something wrong with the structure of the sentence" (ISAC & REIS, 2008, p. 83). For Peter Ludlow, elucidating Chomsky's view, a sentence like 'That's the book that Bill married the woman who illustrated', while perfectly comprehensible, is nevertheless "clearly bad" (2013, p. 06). Maria Teresa Guasti tells us, of three-year-old speakers, that "[a]lthough their language may still not be perfect, they put words *in the correct order*" (GUASTI, 2004, p. 02 – emphasis added). These clearly normative judgements of acceptability seem ubiquitous among language users. Indeed, they constitute the main, perhaps the sole, evidence available to linguists in their primary empirical task of reconstructing speakers' I-Languages;[4] and Chomsky declares that "a theory of language [which] failed to account for these judgements... would plainly be a failure" (1986, p. 37). How, then, can it be claimed that language is not normative?

The key, for anti-prescriptivists, is to distinguish between *what the intuitions express* and *the fact that speakers have these intuitions* (DEVITT, 2008, p. 119; DEVITT & STERELNY, 1989, p. 520-521). The linguist can use these judgements as evidence of how the speaker's I-Language is constructed, without taking their content – including its normative aspect – to be *true*. The normative pull experienced by the speaker is thus regarded as a sort of extra-linguistic gilding; we are trained to police social, class, and political boundaries through our use of language, but this training is not *itself* a part of language, merely one exclusionary use to which language can be put.

The most detailed development of this response is due to Peter Ludlow. If an individual's grammar is the result of the parameters of Universal Grammar (UG) being set during the language acquisition process then he, Ludlow, has the grammar G_{PL} as a result of being in parametric state UG_{PL}. We can then distinguish between the "language narrowly construed"[5] which is *generated* by his grammar – L_{GPL} – and any "other phenomena that we might pre-theoretically take to be linguistic, or part of my 'language' understood loosely speaking" – L_{PL} (LUDLOW, 2013, p. 51). Now, sentences may be well-formed according to L_{GPL}, yet still rejected

[4] For detailed discussion of the controversies over the nature of evidence in linguistics (See DEVITT, 2008, p. 95 *ff.*; LUDLOW, 2013, p. 53-54; 64 *ff.*).
[5] In the sense of Hauser, Chomsky & Fitch (2002) and Fitch, Hauser & Chomsky (2005).

by Ludlow as unacceptable – that is, excluded from L_{PL} – owing to non-linguistic processing limitations, for instance; he gives the example of 'The mouse the cat the dog bit chased ran away'(*ibid.*).

But other extraneous, non-linguistic factors can have a similar effect; if Ludlow rejects 'I ain't got no money', it may be because the sentence in fact violates L_{GPL}, or because he was "inculcated with prescriptive rules… drilled by grammar school teachers not to use "ain't" and "double negatives"" (*ibid.*, p. 51-52).[6] The first of these will be a linguistic reason in the strict sense, the second non-linguistic; and it may prove exceedingly difficult for linguists to discern which is in play. But in either case, there is no reason to think that the normative nature of Ludlow's "surfacey" judgement shows his I-Language – L_{GPL} – to be normative. If the sentence does violate L_{GPL}, then that is an interesting linguistic fact about Peter Ludlow and the parametric state of his mind/brain, with no bearing on any other individual's. If the aversion is a "drilled", inculcated one then it turns out, against appearances, not to be linguistic at all.

2

So stand the two sides of the debate, with little sign of movement. As I will go on to argue, however, a recent 'paralinguistic' phenomenon provides some reason to doubt that the response just outlined is satisfactory.[7] I will first here describe that phenomenon, then go on in the final section of the paper to explain the problem it poses for the anti-prescriptivist position.

The 'Doge' meme, an internet fad that became wildly popular during 2013,[8] even spawning its own currency (HERN, 2014) and consequent cybercrime (SOUPPOURIS, 2013), consists in its canonical form of a photograph of a Shiba Inu dog, surrounded by snippets of interior monologue in brightly-coloured Comic Sans.[9] The meme attracted a surprising amount of attention from linguists, who noted that the snippets had highly distinctive stylistic and grammatical features, to the extent "that doge speak is recognizably doge even when it's not on an image at all" (McCULLOUGH, 2014).[10]

There are two chief kinds of doge phrase. The first is a one-word interjection; usually "wow", "amaze", or "excite". The second consists typically of two words, of which the first is usually "such", "much", "so", "very", or "many"; corpus analysis shows that nearly 40% of all doge phrases begin with one of these five modifiers (NODAR, 2014). A typical doge utterance will combine at least two or three two-word phrases, along with at least one interjection, usually "wow"

[6] As Ludlow goes on to note, this is *not* in fact a double negative.
[7] No doubt other such phenomena could illustrate the same point more or less well. However, I focus on this one for reasons of both clarity and topicality.
[8] https://trends.google.com/trends/explore?date=all&q=doge&hl=en-US; accessed 5th March, 2019.
[9] http://knowyourmeme.com/memes/doge; accessed 5th March, 2019.
[10] McCullough cites one ingenious text-only effort, which begins: 'What light. So breaks. Such east. Very sun. Wow, Juliet.'

(McCULLOUGH, 2014). Hybrid types – *e.g.* "such wow", "very excite" – are also permitted.

The most distinctive grammatical feature occurs in the two-word phrases, and involves "mismatching in the phrasal templates" (GAWNE, 2014), a violation of "selectional restriction" (McCULLOUGH, 2014). That is, the modifier must be one which would *not* usually modify a word of the type which follows it. "So" and "very", for example, in standard usage modify only adjectives; in doge speak they may modify anything *but* an adjective. "Very tasty" and "so delicious" are good English, but poor doge; "very drink" and "so wine" good doge, but poor English. The modified phrase is typically also in its simplest form; hence "amaze" rather than "amazed" or "amazing". Thus, while doge speak looks ungrammatical or grammatically primitive, it is neither; it is "built around a very specific grammar which users wouldn't be able to use unless they had quite a sophisticated grasp of standard English grammar" (CHIVERS, 2014).

3

All very whimsical and entertaining, but what has any of this to do with I-Languages and normativity? Quite a lot, as it happens. To say that doge speak is 'built around a grammar' is, for a prescriptivist, to say that doge speak is composed of *rules*, in the normative, Wittgensteinian sense. That is, failure to abide by the conventions of the Doge meme is not regarded simply as non-standard or idiosyncratic doge speak; it is, precisely, a *failure*. And we can expect other doge users to regard it as such, and to police the rule-following of their interlocutors.

This is, indeed, what we see; users routinely correct others for using constructions that are *too conventionally grammatical*. Linguist Gretchen McCullough (2014) provides a first-hand example:

Friend #1 (posting link): Doge is a rescue dog. Much respect. So noble. Wow.

Friend #2 (commenting): Your dogeing is too coherent. 'Much noble, so respect.'

The most famous case of doge-correction came in December 2013 when U.S. Rep. Steve Stockman (R-Texas) tweeted a doge-style image of primary rival Sen. John Cornyn, featuring the text "wow. kill GOP filibuster. oppose Ted Cruz. support Obamacare funding. don't like Rand Paul." The response was swift and unequivocal: "It's hard to explain the doge meme… but it's definitely not supposed to include full, coherent sentences like 'support Obamacare funding'" (LOGIURATO, 2013); "Aside from 'wow', the words in the photo are just phrases, not doge-isms. Please get it together, Representative" (JONES, 2013); "so correct spelling. not any funny. weak attempt. wow." (ORE, 2013).[11]

[11] Note that the complaints here cannot be construed as merely indicating a pragmatic violation (as in Ludlow's 'mouse/cat/dog' example); there is no difficulty understanding what Stockman, or Friend #1, intended to say in their deviant doge-utterances.

Nor was the reaction simply partisan. Both Stockman and fellow Rep. Thomas Massie (R-Kentucky) – who had tweeted a Shiba Inu image with the caption "Much bipartisanship. Very spending. Wow." – were criticised for "ruining" (McMURRY, 2013), and "killing" (McHugh 2013) the meme by using it for political ends. But while both were accused of crass opportunism, Massie escaped comparatively lightly; "Stockman's tweet was targeted", CBC News explained, "for its flagrant use of grammatically correct phrases – something completely against the spirit of doge" (ORE, 2013).

Of course, Chomsky and his allies are well aware that language-users police each other's rule-following; they simply deny that there is anything *linguistic* about this policing. So why should doge speak present any new problems for them? Recall that such policing was explained by anti-prescriptivists as being "inculcated" and "drilled" into children by parents, teachers, and the general social milieu whose non-linguistic strata and divisions language was being used to enforce. That explanation is *not* obviously available in the case of doge speak; its rules are not learned injunctions, drilled into new users at the time of their socialisation in a particular community. Rather, it looks as though the conventions for using doge speak – its "grammar" in the wide, Wittgensteinian sense – simply are themselves normative. In other words, at least in one admittedly exotic region of human language, normativity is an inherent part of linguistic experience rather than an extraneous accompaniment to it. This shifts the onus of proof considerably; it can no longer be presumptively the case that language, *per se* lacks this feature. To the contrary, the anti-prescriptivist now owes us an explanation of what, if this inherent normativity is not a general feature of language, makes cases such as doge speak special.

Of course, playing "burden tennis" in this way can never be conclusive, and there's an obvious response available here to the I-Language theorist. That is, such a theorist can point out that doge speak *obviously* isn't governed by the principles and parameters of UG. We described it earlier as "paralinguistic"; it uses many of the features of the human language faculty, but in a derivative, "piggybacking" fashion. But it's not *language*; "competence" in doge speak is not part of the speaker's I-Language, and doge speak is not one of the natural human languages that a child can acquire as part of the developmental process of first-language acquisition which the I-Language theory seeks to explain. So the mere fact that the 'grammar' of doge speak is normative does not show that *grammar* properly so-called is.

There are, I think, two ways to construe this move. One is to treat it as demarcating the proper target of linguistic explanation; for some Chomskyans, linguistics is properly concerned only with a distinct subset of the phenomena generally regarded as "linguistic" (the Faculty of Language in the Narrow Sense or FLN). So doge is not a linguistic phenomenon *sensu stricto*, and its normativity shows nothing about purported linguistic normativity. I return to this point below.

The other way of understanding this defense of anti-prescriptivism is to see it as drawing a line between admittedly linguistic phenomena on the basis of their developmental history; though some linguistic behaviours may be learned in a

normatively-laden way, they *are* "special cases" because the "core" linguistic behaviour with which generative linguistics is concerned is acquired by a different, and norm-free, process. But this response, I think, misunderstands the nature of the challenge. What the normativity of doge speak demonstrates is that norms *just do* for whatever reason phenomenologically accompany (at least some) human linguistic (or quasi-, or para-linguistic) behaviour. The developmental history of that behaviour is beside the point, that point being that such norms don't need to "come from" anywhere external in order to form part of our linguistic experience and behaviour; they may simply arise spontaneously as part of the conditions of rule-following in a social context. So there is no reason to think that some extraneous source such as indoctrination is necessary to explain the norms which accompany canonical cases of grammatical speech, when the phenomenologically indistinguishable norms accompanying doge speak can arise without it.

The anti-prescriptivist again seems to have a persuasive answer available here; the phenomenological normativity of language may still be extraneous, but ubiquitous. More precisely, the child learning its language is not taught to attach class or ethnic evaluations to *particular* infractions of its linguistic rules. Rather, it is taught to take *generally* class- or ethnicity-sensitive attitudes towards *any* breach of familiar linguistic conventions; but these attitudes are still not, as such, linguistic. Thus, because doge speak "piggybacks" on normal language, its users' normative attitudes may likewise piggyback on the extraneously-drilled normative attitudes they acquired when they first acquired language. That is, the same extraneous drilling accounts for the normative pull of our I-Language and our doge speak alike.

However, this isn't a solution with which Chomsky and his allies should feel comfortable. Universal Grammar was first invoked to explain the ubiquity of certain structural features across the grammars of all human languages; by the explicitly Cartesian reasoning of the I-Language theorists, a trait universal among humans is likely to be innate to humans (CHOMSKY, 1966). The sheer ubiquity of normative attitudes to language, therefore, creates a defeasible presumption that those attitudes are similarly part of our cognitive and linguistic patrimony. Moreover, though I cannot do more than briefly sketch them here, several related lines of evidence in the literature support this conjecture.

The first line proceeds from the widespread belief that our normative *moral* attitudes – whether or not they possess some further external justification – are just the sort of attitudes we would expect our ancestors to have evolved, given their usefulness in ensuring the cooperative and reciprocal behaviour greatly beneficial to members of a social species like ours (*e.g.* SINGER, 1982; RUSE, 1986; JOYCE, 2006; STREET, 2006; WIELENBERG, 2010; BROSNAN, 2011). Our moral psychology, in other words, has survival value. But our normative *linguistic* attitudes, too – which mark the boundaries of social, ethnic, and national groups – would have had obvious utility in policing complex inter- and intra-group relations, allegiances, and rivalries; a utility which ethnolinguistics suggests they

still possess.[12] So they too are 'the sort of attitudes', if anything is, that we could expect to have inherited from the earliest humans. Indeed, building on Axelrod & Hamilton's (1981) classic analysis of songbird dialects, Daniel Cloud has argued that the complexity of human grammars – and the comparative ease with which children acquire them during the critical developmental window for first-language acquisition – may be adaptations "mostly to make it difficult for adults to learn the language well enough to sound like real natives. At some point in our recent evolutionary history, it might have benefitted us, as it does birds, to be able to quickly tell the difference between genuine members of our own tribe and interlopers" (2015, p. 110-111). And like moral norms, it is perfectly consistent where linguistic norms are concerned to think both that we are drilled in them at our mother's knee, and that we have evolved to hold them. If it is plausible that our normative moral attitudes evolved, then it looks equally plausible that normative attitudes to language have.

Anti-prescriptivists acknowledge that normative attitudes are part of the psychological make-up of language-speakers, but hold that they are drilled and inculcated into us from external sources. What the second line of argument suggests is that, as long as there is survival value to being able to learn them quickly and easily, such psychological traits may well *originate* externally in this fashion, but become progressively assimilated into the genome over many generations via what is known as a 'Baldwin Effect' (BALDWIN, 1896). More precisely, while such traits may be phylogenetically external, at least to begin with, they are ontogenetically internalised in modern humans. In recent years, evolution of the UG via such a mechanism has been independently proposed by several theorists to explain various features of human language (*e.g.* PINKER & BLOOM 1990; DOR & JABLONKA, 2000; JABLONKA & LAMB, 2005; ANDERSON, 2008; SZATHMÁRY, 2010; ANDERSON, 2011; GLACKIN, 2011; ANDERSON, 2013; GLACKIN, 2018). Extensive computational modelling confirms the plausibility of these hypotheses (*e.g.* STEELS, 2011; SUZUKI & ARITA, 2013; AZUMAKIGATO, SUZUKI & ARITA, 2013). Again, if it is plausible that the ubiquitous features of human language which form the UG evolved by this kind of mechanism, it is at least as plausible that the ubiquitous normative attitudes which humans hold towards language did so too.

The first two lines of argument suggest that normative attitudes towards language could have become 'hardwired' in the human genome. The third line suggests that such hardwiring is not actually necessary for the overall point here, once we have come to regard such attitudes as a ubiquitous feature of human linguistic behaviour. Contemporary biological thinking is increasingly moving away from a 'genocentric', reductionist understanding of evolutionary processes, towards one that stresses genes' conceptual dependence on environmental conditions, and the cross-generational transmission of non-genetic information and resources, including human culture (*e.g.* OYAMA, 1985; GODFREY-SMITH, 1996; JABLONKA

[12] *E.g.* "Once a nation or tribe splits in two, each with its own political organization, the two groups will seize on linguistic features as tokens of self-identification. A handful of lexemes and/or pronouns can be sufficient. The dialects of two new nations or tribes may well be fully intelligible, the important political thing being to take care to use certain words and to avoid others" (DIXON, 1997, p. 58). It is from just such a biblical story that the term 'shibboleth' is derived.

& LAMB, 2005; STERELNY, 2012). Chomsky himself (2010) has expressed support for this move, and argued for its applicability to the human language faculty.

Accordingly, we can regard our drilling and inculcation in normative linguistic attitudes as part of our cognitive inheritance, which has shaped both our minds and our linguistic practices, *whether or not any genetic hardwiring took place*.[13] Since vertical genetic and cultural transmission are equally legitimate and biologically significant modes of intergenerational information flow, there is no principled argument for excluding the cultural part of our linguistic inheritance – if that is, as Chomsky and his followers claim, "all" that our learned normative attitudes to linguistic practice represent – from biolanguage.

We arrive, then, at the view that normative attitudes towards language, which are as robustly ubiquitous a feature of human language-use as any other, are therefore as much and as central a part of our linguistic inheritance as any other. That they are typically taught to us externally, unlike the supposedly "innate" and "automatic" operations of the UG, makes them no less natural or normal a feature of linguistic development; human language, both in its practice and its phenomenology, looks – in Wilfrid Sellars' phrase – thoroughly "fraught with ought" (MILLIKAN, 2005, p. 79).

There is one remaining move available, however, to anti-prescriptivists, which I flagged above; that is to return to the distinction between 'language in the narrow sense' (FLN) and 'language in the broad sense' (FLB) as Ludlow did previously (§1; LUDLOW, 2013, p. 51). That is, I-Language theorists might grant all of the foregoing, but nevertheless insist that these universal, inherited attitudes are still not, strictly speaking, *part of language*. Fitch, Hauser, and Chomsky define FLB as "all of the many mechanisms involved in speech and language, regardless of their overlap with other cognitive domains or with other species" (2005, p. 179-180) and FLN as whatever "subset of the mechanisms of FLB is both unique to humans, and to language itself" (*ibid.*, p. 180). This subset, they hypothesise, is limited to the internal computational system which handles syntactical recursion (*ibid.*, p. 203 *ff.*; HAUSER, CHOMSKY & FITCH, 2002, p. 1571), and would thus

[13] There is a further point to be made here, though it does not form part of the main thread of the argument. As Cloud has shown, complex informational resources, if they are to be reproduced across generations without succumbing to "error catastrophe" due to the accumulation of mutations (EIGEN, 1992, p. 20), must be accompanied by corrective mechanisms; since imitation has a particularly poor fidelity of replication, an informational resource substantially passed on in this fashion – as is typical of human culture (e.g. GRIMM, 2000; STERELNY, 2012) – is crucially dependent upon the awareness of the imitated party that they are being imitated, and their willingness to provide feedback by correcting failures of imitation. This explains a large part of humans' capacity for culture, which is not shared by chimpanzees despite their cognitive resources; "Humans imitate a lot, and humans correct one another's mistakes a lot. Chimpanzees and other apes don't imitate very much; they mostly emulate, and they very seldom correct one another's mistakes. This… must be at least partly because the fidelity of their imitations would be too low to avoid error catastrophe, and make imitating a good idea for the typical individual in the typical population" (CLOUD, 2015, p. 131-132). This insight, speculative though it is, provides further support for the first line of argument traced above; insofar as the conventions of human language form a complex informational resource to be transmitted across generations, there will be selective pressure for the tendency to correct linguistic "errors" – failures to replicate the convention accurately – in others, as well as to accept such correction from others, and modify linguistic behaviours accordingly.

exclude the normative attitudes we are interested in. That exclusion looks correct, as it goes; besides the possibility that these attitudes use the same cognitive apparatus as our moral attitudes, there is some evidence for precursors to normativity in animal communication (HAUSBERGER *et al.*, 2008; LACHLAN, 2008).

But this isn't enough to show that human language is not, *per se*, normative, or that those normative attitudes are not part of the province of linguistics. In fact, Fitch, Hauser, and Chomsky repeatedly stress that the mechanisms making up FLN are "neither the only, nor necessarily the most, interesting problems for biolinguistic research" (2005, p. 181). Again; "we don't suggest that only phenomena in FLN are worthy of study" (*ibid.*, p. 203). And most pertinently; "(w)e doubt that future researchers will need to make a point of distinguishing FLN from FLB at every mention of the word 'language', as we have done here" (*ibid.*, p. 205). So the Ludlovian move of placing normative attitudes and the processes which produce them outside of language "narrowly construed" doesn't thereby establish that they are only "pre-theoretically" to be regarded as linguistic.

There's a deeper problem lurking here for the FLN/FLB distinction, too. As I have elsewhere (GLACKIN, 2018, p. 174) pointed out, the "first motivation" (BERWICK & CHOMSKY, 2016, p. 11) for the introduction of the FLN as the true object of linguistic study, and the accompanying "minimalist" view of linguistics, is to reduce the explanandum for a saltationist theory of language's evolution. Since Chomsky and his followers regard a gradual evolutionary process for the language faculty as implausible, that faculty must be such that it could be achieved by a sudden process instead; as small and as un-complex as possible. But if this evolutionary reasoning is flawed, as I have argued, then the motivation for the FLN/FLB distinction disappears; there is no good reason to accept the FLN as the proper and unique subject matter of generative linguistics – and to thereby exclude our normative attitudes to language from the province of the linguistic – unless one adopts a series of controversial assumptions about the nature of evolutionary theory.

In short, what the Doge meme shows us is that the normative attitudes we adopt towards grammatical rules cannot simply be a set of learned particular prescriptions;[14] rather, whether it is learned or innate, they must result from a generalised normative attitude towards such rules. And a general, ubiquitous normative attitude towards linguistic rules, whether learned or innate, could only arbitrarily be excluded from the province of language, and the subject-matter of human biolinguistics. Chomsky and his followers are certainly entitled to hold that such norms have nothing to do with *their* research project. They are right to point out, too, that much of the norms' interest is sociological or political; but humans are, of course, social and political animals, and any comprehensive biolinguistics must take those facets into account. Prescriptive norms are a real and ubiquitous feature of language, and a real and legitimate object of study for linguists.

[14] To avoid any ambiguity; whether the grammatical rules themselves are learned or innate in any particular case, my claim here is that *our normative attitudes towards them* are not learned as particular prescriptions.

References

ANDERSON, Stephen R. The logical structure of linguistic theory. *Language*, v. 84, p. 795-814, 2008.

_____. The role of evolution in shaping the human language faculty. In: M. TALLERMAN, Maggie; GIBSON, Kathleen R. (Eds.). *The Oxford handbook of language evolution*. Oxford: Oxford University Press, 2011. p. 361-369.

_____. What is special about the human language faculty and how did it get that way? In: BOTHA, Rudolf; EVERAERT, Martin (Eds.). *The evolutionary emergence of language*: Evidence and inference. Oxford: Oxford University Press, 2013. p. 18-41.

AXELROD, Robert; HAMILTON, William D. The evolution of cooperation. *Science*, v. 211, p. 1390-1396, 1981.

AZUMAKIGATO, Tsubasa; SUZUKI, Reiji; ARITA, Takaya. Cyclic behavior in gene-culture coevolution mediated by phenotypic plasticity in language. In: LIÒ, Pietro et al (Eds.). *Advances in artificial life, ECAL 2013*: Proceedings of the twelfth European conference on the synthesis and simulation of living systems. Cambridge, MA: MIT Press, 2013. p. 617-624.

BALDWIN, James M. A new factor in evolution. *The American Naturalist*, v. 30, p. 536-553, 1896.

BERWICK, Robert C.; CHOMSKY, Noam. *Why only us*: Language and evolution. Cambridge, MA: MIT Press, 2016.

BROSNAN, Kevin. Do the evolutionary origins of our moral beliefs undermine moral knowledge? *Biology and Philosophy*, v. 26, p. 51-64, 2011.

CHIVERS, Tom. Doge: Such grammar. Very rules. Most linguistic. Wow. *The Daily Telegraph Blogs*, February 18, 2014. Available at: http://blogs.telegraph.co.uk/news/tomchiversscience/100260120/doge-such-grammar-very-rules-most-linguistics-wow/. Accessed in: August 13, 2014.

CHOMSKY, N. *Cartesian linguistics*: a chapter in the history of rationalist thought. New York: Harper & Row, 1966.

CHOMSKY, Noam. *Knowledge of language*: Its nature, origin, and use. New York: Praeger, 1986.

_____. Language and problems of knowledge. In: MARTINICH, A. P. (Ed.). *The philosophy of language*. 5.ed. New York: Oxford University Press, 2010. p. 675-692.

_____. Explaining language use. *Philosophical Topics*, v. 20, p. 205-231, 1992a.

_____. Language and interpretation. In: EARMAN, John (Ed.). *Inference, explanation and other frustrations*. Berkeley: University of California Press, 1992b. p. 99-128.

CHOMSKY, Noam. Mental constructions and social reality. In: REULAND, Eric; ABRAHAM, Werner (Eds.). *Knowledge and language - vol. 1*: From Orwell's problem to Plato's problem. Dordrecht: Kluwer Academic, 1993. p. 29-58.

_____. Language and nature. *Mind*, 104, p. 1-61, 1995.

_____. *New horizons in the study of language and mind*. Cambridge: Cambridge University Press, 2000.

_____. Some simple evo devo theses: How true might they be for language? In: LARSON, Richard K.; DÉPREZ, Viviane; YAMAKIDO, Hiroko (Eds.). *The evolution of language*: biolinguistic perspectives. Cambridge: Cambridge University Press, 2010. p. 45-62.

_____. Interview with Noam Chomsky. In: LUDLOW, Peter. *The philosophy of generative linguistics*. Oxford: Oxford University Press, 2013. p. 174-191.

CLOUD, Daniel. *The domestication of language*: cultural evolution and the uniqueness of the human animal. New York: Columbia University Press, 2015.

DEVITT, Michael. *Ignorance of language*. Oxford: Oxford University Press, 2008.

DEVITT, Michael; STERELNY, Kim. What's wrong with 'the right view'. In: TOMBERLIN, James (Ed.). *Philosophical perspectives, 3*: philosophy of mind and action theory. Atascadero, CA: Ridgeview, 1989. p. 497-531.

DIXON, R. M. W. *The rise and fall of languages*. Cambridge: Cambridge University Press, 1997.

DOR, Daniel; JABLONKA, Eva. From cultural selection to genetic selection: a framework for the evolution of language. *Selection*, v. 1, p. 33-55, 2000.

EIGEN, Manfred. *Steps towards life*: a perspective on evolution. Oxford: Oxford University Press, 1992.

FITCH, W. T.; HAUSER, Marc D.; CHOMSKY, Noam. The evolution of the language faculty: clarifications and implications. *Cognition*, v. 97, p. 179-210, 2005.

GAWNE, Lauren. Doge - meme watch. *Superlinguo*, November 20, 2013. Available at: http://www.superlinguo.com/post/67501611729/doge-meme-watch. Accessed in: March 5, 2019.

GLACKIN, Shane N. Universal grammar and the Baldwin effect: a hypothesis and some philosophical consequences. *Biology and Philosophy*, v. 26, n. 2, p. 201-222, 2011.

_____. Against Thatcherite linguistics: rule-following, speech communities, and biolanguage. *The Southern Journal of Philosophy*, v. 56, n. 2, p. 163-192, 2018.

GODFREY-SMITH, Peter. *Complexity and the function of mind in nature*. Cambridge: Cambridge University Press, 1996.

GRIMM, Linda. Apprentice flint-knapping: relating material culture and social practice in the upper paleolithic. In: SOFAER DEREVENSKI, Joanna (Ed.). *Children and material culture*. New York: Routledge, 2000. p. 53-71.

GUASTI, Maria Teresa. *Language acquisition*: the growth of grammar. Cambridge, MA: MIT Press, 2004.

HAUSBERGER, Martine et al. Contextual sensitivity and bird song: a basis for social life. In: KIMBROUGH OLLER, D.; GRIEBEL, Ulrike (Eds.), *Evolution of communicative flexibility*: complexity, creativity, and adaptability in human and animal communication. London: MIT Press, 2008. p. 121-138.

HAUSER, Marc D.; CHOMSKY, Noam; FITCH, W. T. The language faculty: what is it, who has it, and how did it evolve? *Science*, v. 298, p. 1569-1579, 2002.

HERN, Alex. What is doge? *The Guardian*, February 18, 2014. Available at: http://www.theguardian.com/technology/2014/feb/18/doge-such-questions-very-answered. Accessed in: March 5, 2019.

ISAC, Daniela; REISS, Charles. *I-language*. Oxford: Oxford University Press, 2008.

JABLONKA, Eva; LAMB, Marion J. *Evolution in four dimensions*: genetic, epigenetic, behavioural, and symbolic variation in the history of life. Cambridge, MA: MIT Press, 2005.

JONES, Allie. Cool politicians use doge memes to prove they really 'get' twitter. *The Wire*, December 23, 2013. Available at: http://www.thewire.com/politics/2013/12/politicians-trying-be-cool-twitter/356429/. Accessed in: March 5, 2019.

JOYCE, Richard. *The evolution of morality*. Cambridge, MA: MIT Press, 2006.

KRIPKE, Saul. *Wittgenstein on rules and private language*. Oxford: Blackwell, 1982.

LACHLAN, Robert F. The evolution of flexibility in bird song. In: KIMBROUGH OLLER, D.; GRIEBEL, Ulrike (Eds.). *Evolution of communicative flexibility*: complexity, creativity, and adaptability in human and animal communication. London: MIT Press, 2008. p. 305-326.

LOGIURATO, Brett. Congress has finally discovered 'doge', and it's going about as badly as you would expect. *Business Insider*, December 23, 2013. Available at: http://www.businessinsider.com/doge-memes-congress-steve-stockman-massie-2013-12#ixzz39pkTBuwb. Accessed in: 5 March, 2019.

LUDLOW, Peter. *The philosophy of generative linguistics*. Oxford: Oxford University Press, 2013.

McCULLOUGH, Gretchen. A linguist explains the grammar of doge. Wow. *The Toast*, February 06, 2014. Available at: http://the-toast.net/2014/02/06/linguist-explains-grammar-doge-wow/. Accessed in: 4 March, 2019.

McHUGH, Molly. The life and death of doge, 2013's greatest meme. *The Daily Dot*, December 23, 2013. Available at: http://www.dailydot.com/lol/doge-is-dead-long-live-doge/. Accessed in: 5 March, 2019

McMURRY, Evan. Rep. Steve Stockman ruins doge for the rest of us. *Mediaite*, December 23, 2013. Available at: http://www.mediaite.com/online/rep-steve-stockman-ruins-doge-for-the-rest-of-us/. Accessed in: 5 March, 2019.

McNALLY, Thomas; McNALLY, Sinéad. Chomsky and Wittgenstein on linguistic competence. *Nordic Wittgenstein Review*, v. 1, p. 131-154, 2012.

MILLIKAN, Ruth G. Pushmi-pullyu representations. In: TOMBERLIN, James (Ed.). *AI, connectionism, and philosophical psychology*. Atascadero, CA: Ridgeview, 1995.

_____. In defense of public language. In: ANTONY, Norbert; HORNSTEIN, Louise (Eds.). *Chomsky and his critics*. Oxford: Blackwell, 2003. p. 215-37.

_____. The son and the daughter: on Sellars, Brandom, and Millikan. *Pragmatics and Cognition*, v. 13, n. 1, p. 59-71, 2005.

NODAR, Leah. The curious linguistics of the doge in the internet. *The League of Nerds*, March 04, 2014. Available at: http://www.asktheleagueofnerds.com/doge/. Accessed in: 5 March, 2019.

ORE, Jonathan. U.S. republican doge tweets 'wow' internet meme fans. *CBC News Community Blogs*, December 24, 2013. Available at: http://www.cbc.ca/newsblogs/yourcommunity/2013/12/us-republican-doge-tweets-wow-internet-meme-fans.html. Accessed in: 5 March, 2019.

OYAMA, Susan. *The ontogeny of information.* Cambridge: Cambridge University Press, 1985.

PINKER, Steven; BLOOM, Paul. Natural language and natural selection. *Behavioral and Brain Sciences*, v. 13, p. 707-784, 1990.

RUSE, Michael. *Taking Darwin seriously*. New York: Blackwell, 1986.

SINGER, Peter. Ethics and sociobiology. *Philosophy and Public Affairs*, v. 11, p. 40-64, 1982.

SOUPPOURIS, Aaron. Millions of dogecoin stolen in Christmas hack. *The Verge*, December 26, 2013. Available at: http://www.theverge.com/2013/12/26/5244604/millions-of-dogecoin-stolen-in-christmas-hack Accessed in: 5 March, 2019.

STEELS, Luc. Modelling the cultural evolution of language. *Physics of Life Reviews*, v. 8, n. 4, p. 339-356, 2011.

STEINER, M. Kripke on logicism, Wittgenstein, and *de re* beliefs about numbers. In: BERGER, Alan (Ed.). *Saul Kripke*. Cambridge: Cambridge University Press, 2011. p. 160-176.

STERELNY, Kim. *The evolved apprentice*: how evolution made humans unique. Cambridge, MA: MIT Press, 2012.

STREET, Sharon. A Darwinian dilemma for realist theories of value. *Philosophical Studies*, v. 127, p. 109-166, 2006.

SUZUKI, Reiji; ARITA, Takaya. A simple computational model of the evolution of a communicative trait and its phenotypic plasticity. *Journal of Theoretical Biology*, v. 330, p. 37-44, 2013.

SZATHMÁRY, Eörs. Evolution of language as one of the major evolutionary transitions. In: NOLFI, Stefano; MIROLLI, Marco (Eds.). *Evolution of communication and language in embodied agents*. Berlin: Springer Verlag, 2010. p. 37-53.

WALLACE, David F. Tense present: Democracy, English, and the wars over usage. *Harper's Magazine*, v. 302, p. 39-58, April, 2001.

WIELENBERG, Erik J. On the evolutionary debunking of morality. Ethics, v. 120, p. 441-464, 2010.

WITTGENSTEIN, Ludwig. *Philosophical investigations*. Oxford: Basil Blackwell, 1953.

Recebido em: 06-03-2019

Aceito para publicação em: 30-07-19

NEO-MECHANISTIC EXPLANATORY INTEGRATION FOR COGNITIVE SCIENCE: THE PROBLEM OF REDUCTION REMAINS

INTEGRAÇÃO EXPLANATÓRIA NEOMECANICISTA PARA A CIÊNCIA COGNITIVA: O PROBLEMA DA REDUÇÃO PERMANECE

DIEGO AZEVEDO LEITE[1]
Università Degli Studi di Trento – Italy
diego.azevedo.leite@gmail.com

ABSTRACT: One of the central aims of the neo-mechanistic framework for the neural and cognitive sciences is to construct a pluralistic integration of scientific explanations, allowing for a weak explanatory autonomy of higher-level sciences, such as cognitive science. This integration involves understanding human cognition as information processing occurring in multi-level human neuro-cognitive mechanisms, explained by multi-level neuro-cognitive models. Strong explanatory neuro-cognitive reduction, however, poses a significant challenge to this pluralist ambition and the weak autonomy of cognitive science derived therefrom. Based on research in current molecular and cellular neuroscience, the framework holds that the best strategy for integrating human neuro-cognitive theories is through direct reductive explanations based on molecular and cellular neural processes. It is my aim to investigate whether the neo-mechanistic framework can meet the challenge. I argue that leading neo-mechanists offer some significant replies; however, they are not able yet to completely remove strong explanatory reductionism from their own framework.

KEYWORDS: Integration of cognitive science. Neuro-cognitive integration. Neo-mechanistic explanation. Explanatory pluralism. Explanatory integration.

RESUMO: *Uma das finalidades centrais da abordagem teórica neomecanicista para as ciências neural e cognitiva é a construção de uma integração pluralística de explicações científicas, permitindo uma autonomia explanatória fraca das ciências de mais alto nível, como a ciência cognitiva. Essa integração envolve a compreensão da cognição humana como processamento de informação ocorrendo em mecanismos neurocognitivos de múltiplos níveis, explicados por modelos neurocognitivos de múltiplos níveis. A redução explanatória neurocognitiva forte, no entanto, apresenta um desafio significativo para esta ambição pluralista e a autonomia fraca da ciência cognitiva dela derivada. Baseada em pesquisas na área atual da neurociência molecular e celular, essa abordagem teórica sustenta que a melhor estratégia para integrar teorias da neurocognição humana é através de explicações redutivas diretas baseadas em processos neurais moleculares e celulares. O meu objetivo é investigar se a estrutura teórica neomecanicista pode superar esse desafio. Eu argumento que os principais neomecanicistas oferecem algumas respostas significativas; porém, eles ainda não são capazes de remover completamente o reducionismo explanatório forte da sua própria estrutura teórica.*

PALAVRAS-CHAVE: *Integração da ciência cognitiva. Integração neurocognitiva. Explicação neomecanicista. Pluralismo explanatório. Integração explanatória.*

[1] PhD in Cognitive Science from the University of Trento, Italy (2018), in the line of research on Philosophy of Cognitive Science.

INTRODUCTION

The new *mechanistic theory of scientific explanation*, articulated in the end of the 20th century and the beginning of the 21st century, is one of the most important and influential contemporary theories of scientific explanation (BECHTEL; RICHARDSON, 1993/2010; BECHTEL; ABRAHAMSEN, 2005; MACHAMER; DARDEN; CRAVER, 2000; CRAVER; TABERY, 2015; GLENNAN; ILLARI, 2018). This theory is applied especially to the biological sciences, including cognitive science.

This neo-mechanistic framework applied to cognitive science is advocated by many contemporary influential authors (e.g. BECHTEL, 1994, 2007, 2008, 2009a, 2009b, 2009c, 2010, 2017; BECHTEL; WRIGHT, 2009; BOONE; PICCININI, 2016; CRAVER, 2002, 2007; KAPLAN, 2017; MILKOWSKI, 2016; PICCININI; BAHAR, 2013; PICCININI; CRAVER, 2011; PICCININI, 2007, 2012, 2015; THAGARD, 2006, 2009, 2018; WRIGHT; BECHTEL, 2007; ZEDNIK, 2018). Its central idea is that any human cognitive process is a kind of neural process, understood in terms of cognitive information processing and cognitive representation. Such processes can be decomposed and localized in brains as parts of a multilevel neural-biological mechanism. As a result, the framework suggests a path for a general 'pluralistic integration' of neuro-cognitive theories, allowing at the same time for some kind of weak autonomy[2] of scientific explanations in cognitive science.

However, central theoretical aspects of the neo-mechanistic framework for cognitive science remain highly controversial (cf. LEITE, 2018). Particularly, one of the major problems is about whether the neo-mechanistic framework can indeed provide the pluralist integration and weak explanatory autonomy of cognitive science it promises. One of the main obstacles comes from contemporary neuro-cognitive explanatory reductive approaches to the issue. John Bickle can be regarded as one of the most prominent representatives of contemporary neuroscientific explanatory reductionist ambitions (BICKLE, 2003, 2006, 2008, 2012, 2015, 2016). He contends that one should not see reduction in a negative manner in science, nor give up on the scientific general reductionist project, since reductionism can improve scientific disciplines mainly through theoretical unification and elimination of explanatory redundancy (BICKLE, 2003). He argues

[2] Bechtel has an article, published in 2007, titled "Reducing psychology while maintaining its autonomy via mechanistic explanations". In this work, he writes "I will argue in subsequent sections that the reductions achieved through mechanistic explanations are in fact compatible with a robust sense of autonomy for psychology" (2007, p. 174). Moreover, in his famous work of 2008, Bechtel writes: "Traditionally, arguments for the autonomy of psychology have appealed to the claimed multiple realizability of mental phenomena. I contend that there is no evidence for the sort of multiple realizability claimed, but that recognizing this does not undercut the case for autonomous inquiries at higher levels of organization" (p. xi). Craver (2007) also writes: "The different fields that contribute to the mosaic unity of neuroscience are autonomous in that they have different central problems, use different techniques, have different theoretical vocabularies, and make different background assumptions; they are unified because each provides constraints on a mechanistic explanation." (p. 231). And Boone and Piccinini (2016) also discuss traditional "strong autonomy" (related especially with Jerry Fodor), the kind of autonomy they oppose. Given this, I use in this paper the technical term 'explanatory weak autonomy' to clarify *what kind of autonomy* I am discussing. It is not the traditional strong sense, but there is still a sense in which the neo-mechanistic integration includes some weak explanatory autonomy.

for a general explanatory neuroscientific reductive hypothesis on the human neuro-cognitive relationship, based on what neuroscientists are currently doing in the field of molecular and cellular neuroscience.

In this paper, my aim is to investigate whether the neo-mechanistic framework is able to provide a consistent defence of its 'pluralistic integration' in cognitive science in spite of the challenge presented by the neuroscientific neo-reductionist approach. To achieve this goal, firstly, I provide a more detailed characterization of the 21st century mechanistic framework applied to human cognition in cognitive science and discuss its application to the process of memory consolidation (section 1). After this, I characterize the strong neuro-cognitive reductionist approach and its application to the same process of memory consolidation (section 2). The contrast of both approaches shows the challenges the reductionist position brings to the mechanistic one. After the discussion of the divergent points of the two approaches, I analyse the replies offered by the proponents of the neo-mechanistic framework (section 3) and the counter-replies offered by the neuro-cognitive reductionist position (section 4). Based on these analyses, I discuss to what extent the neo-mechanistic framework is successful in its defense of weak explanatory autonomy of cognitive science (section 5). I argue that the framework ultimately incorporates a strong explanatory reductionist position. As a result, any kind of explanatory autonomy for cognitive science is also eliminated.

1 The neo-mechanistic framework for cognitive science

Despite of the many theoretical and terminological differences between the accounts presented by leading neo-mechanists in the field of cognitive science, there are some important common basic points.

The particular application of the general neo-mechanistic account to cognitive science generates a particular theory about scientific activity in the field, which can be called the *mechanistic theory of scientific explanations in cognitive science*. At the same time, this application also generates a theory about the nature of human cognition[3] and of the human neuro-cognitive relationship, which are the most important objects of investigation and explanation in cognitive science. This second theory can be called the *mechanistic theory of human cognition*. These two theories are, evidently, strictly related: the second is concerned with the *explanandum*, and the first with the *explanans*. Together, they provide the neo-mechanistic general framework for investigating human cognition in cognitive science.

[3] Neo-mechanists do not claim that there is something peculiar about human cognition. They assume there is nothing distinctive about it, but this is controversial. Leading cognitive scientists would argue that human cognition has many particular features not found in other cognitive natural organisms, and that aspects of the cognition in these organisms cannot be simply extrapolated to human cognition (cf. BRUNER, 1990; VON ECKARDT, 1993). Thus, I am not using this term because I think neo-mechanists use it, but rather because I think it is more precise and shows more clearly the focus and scope of my discussion.

Thus, the mechanistic theory of human cognition presents an account of what needs to be explained, namely, the human neurocognitive biological mechanism. A biological complex 'mechanism' is normally characterized as "a structure performing a function in virtue of its component parts, component operations, and their organization" (BECHTEL, 2008, p. 13).[4] The core idea is that biological complex mechanisms are systems behaving in a given environment. The general behavior of the whole mechanism is a result of the specific organization of the components and their interactions. According to the mechanistic theory of human cognition, human neuro-cognitive processes are, roughly, information processes performed by neural mechanisms, which represent physical information (cf. BECHTEL, 2008, 2009b; CRAVER, 2007; PICCININI, 2012; THAGARD, 2006; ZEDNIK, 2018). These processes and the mechanisms that perform them can be decomposed in subparts, and these subparts further decomposed. As a result, there can be multiple levels of mechanistic composition in a human neuro-cognitive mechanism. Furthermore, relevant autonomous processes of causation happen in all these different levels (BECHTEL, 2017; CRAVER; BECHTEL, 2007). According to the mechanistic theory of scientific explanations in cognitive science, all these levels and causal processes, in spite of being autonomous, can be related in a pluralistic mechanistic explanation, where the relevant scientific theories are integrated.

One of the clearest examples of an application of the neo-mechanistic framework to a cognitive process is in the domain of memory (cf. BECHTEL, 2008, 2009a; CRAVER, 2007). One important phenomenon related to memory is memory consolidation. Roughly put, this is the phenomenon of transforming short-term memories into long-term memories, what permits the organism to remember important events for a longer period of time and modify its behavior accordingly. To explain this phenomenon, all the relevant regions in the brain responsible for the functions that compose the neuro-cognitive mechanism of memory consolidation, including all relevant mechanistic levels of decomposition, must be identified through the process of localization, i.e. all the particular component parts and component operations of the whole mechanism must be determined.[5] Finally, the causal processes and causal interactions within the functions of the mechanism need also to be understood, i.e. the general organization of the mechanism, and all the different mechanistic levels relevant for the explanation of the phenomenon must be related.

[4] See also the formulations in Craver (2007) and Glennan and Illari (2018).

[5] Firstly, there is the neural systems level which includes for instance "the hippocampus" and its "neuro-architecture" (BECHTEL, 2009a, p. 16, 22). This large neural network (that includes the hippocampus and other particular areas in the brain) is the whole mechanism; and one of the functions that this whole mechanism performs is the phenomenon of memory consolidation, which is the *explanandum* target. The explanation can go further then to a second level, when it decomposes the large neural system into particular sub-neural systems, i.e. a larger neural network that involves larger regions in the brain into smaller neural networks, which are more localized in particular regions. The explanation can go further to another level of decomposition: the inter-cellular level. At this particular level, the components of a particular neural network need to be correctly understood. Finally, the explanation can go to an even lower mechanistic level: the intra-cellular and molecular level. At this level, the description is in terms of the activity of relevant proteins, molecules and ions (cf. BECHTEL, 2009a, p. 18).

2 The strong explanatory neuro-cognitive reductionism

In John Bickle's view, fields such as cognitive neuroscience and cognitive science do not provide the ultimate most complete explanations of human cognition. Instead, he argues that there is an area of current neuroscience, the mainstream of the discipline, which can provide such explanations and is at the same time ruthlessly reductive in spirit: molecular and cellular neuroscience. In his view, at the molecular and cellular lower-level of neural activity very much is already known and it is false that "lower-level neuroscience cannot explain cognition and complex behavior directly." (BICKLE, 2006, p. 411; cf. p. 414). Molecular pathways inside individual neural cells can be linked with cognition and these links "are reductions" (BICKLE, 2008, p. 37).

To accomplish a reduction in Bickle's sense some important steps are needed. The first one is to intervene using molecular genetic methodology into the genome of animals, usually mice. The aim is "to increase or decrease in vivo gene expression and subsequent protein synthesis of intracellular signaling molecules known to be components of pathways that induce and maintain activity-driven synaptic plasticity" (BICKLE, 2012, p. 100). The second step is to measure the effects of the intervention in the behavior of the organism under controlled experimental conditions. The genetically modified animals perform a variety of behavioral tasks so that their specific 'cognitive functions' can be measured. Their behaviors are contrasted with the control animals who are not genetically manipulated. The significant behavioral differences are then understood as the result of the genetic manipulations: the differences in the genetic mechanisms of protein production is the most relevant direct causal factor for explaining the modification in the cognitive function that ultimately produces the behavioral differences (BICKLE, 2012, p. 100).[6]

According to Bickle (2012, p. 101), some scientific experiments already present evidence for establishing the connection between a molecular and cellular mechanism and a particular behavior that indicates a cognitive function. The most clear and detailed example discussed by Bickle is also related to memory, which provides the field of molecular and cellular neuroscience with its "most impressive

[6] The field of molecular and cellular neuroscience has as its central characteristic, thus, the "application of transgenic techniques from molecular genetics into neuroscience. These features allow experimenters to mutate any cloned gene in living, behaving mammals, and thereby manipulate key proteins in intracellular signaling pathways" (BICKLE, 2015, p. 305). These techniques increased the scientific capacity of "manipulation specificity and control" in neuroscience, and generally they "enable a clear picture of not only which neurons have been manipulated but also the specific intracellular signaling pathways affected in those neurons" (BICKLE, 2015, p. 306). In this way, when manipulations of the organisms are successfully made and they produce significant changes in the related behaviors (which can be measured), one can claim that the neural causal-mechanism explains those behaviors, and thereby "the cognitive functions those behaviors are taken to indicate", i.e. "cognitive behaviors" (BICKLE, 2015, p. 305, 306). For the neo-reductionist, research in this area, accordingly, "tests, directly and experimentally, causal-mechanistic hypotheses that purport to explain cognition" (BICKLE, 2015, p. 306). Moreover, in Bickle's view, the development and application of these techniques of engineering genetically mutated mammals, together with the development and applications of the relatively recent technique called 'optogenetics', in order to explain cognition in neuroscience (especially in cellular/molecular neurobiology and behavioral neuroscience) can be considered genuine scientific "revolutions" (2016, p. 1, 2).

achievements" (BICKLE, 2008, p. 36). In one scientific experiment a mouse in which the protein (transcription factor) CREB was 'knocked-out' had intact short-term memory on many rodent memory tasks, while in the long-term memory versions of these tasks there was a great decrease in the memory capacity in comparison with the control. In another experiment CREB was increased in a small population of neurons in a manipulated mouse. This led to an increase of memory consolidation, measured by fear conditioning behavior in the modified mouse. Since CREB has been traditionally considered to be implicated in the induction of late long-term potentiation (L-LTP – a form of neural activity that can last hours, days or weeks which increases neurotransmission efficacy at individual chemical synapses), ultimately is the presence of CREB that is doing, arguably, the bottom-level most central causal work. In fact CREB is part of a particular molecular mechanism that involves cAMP, PKA and CREB, which leads to L-LTP. Bickle claims that blocking any step of this mechanistic process "virtually eradicates memory consolidation, while enhancing steps can lead to faster and stronger consolidation" (2008, p. 38). While these particular experiments are performed mostly in mice, the fundamental hypothesis is that cognition in general can be explained in this way, including human cognition.

The central idea is that this dynamics at the molecular level is actually what *produces* the cognitive processes called long-term memory and memory consolidation. What is being asserted is that there is already sufficient empirical experimental evidence to establish a "causal connection between a proposed cellular or molecular mechanism and a complex, system-level cognitive phenomenon" (BICKLE, 2012, p. 102; cf. SILVA; BICKLE, 2009; SILVA; LANDRETH; BICKLE, 2014).

In order to establish these causal connections, though, Bickle admits that "higher-level scientific investigations" are necessary (BICKLE, 2012, p. 103). This is because precise knowledge about how the whole system behaves is important in order to correlate the "proposed molecular mechanism and the system's behavior we use to indicate the occurrence of a specific cognitive function" (BICKLE, 2012, p. 103). Behavioral experiments at a higher-level also help to establish the "theoretical plausibility of the proposed molecular mechanism for that cognitive phenomenon" (BICKLE, 2012, p. 103-104). Moreover, cognitive neuroscience and its mechanistic goals of decomposition and localization is also important, since it is necessary to identify the most relevant types of neural activity. Consequently, cognitive neuroscience and cognitive science are not dismissed from the relevant and necessary scientific activity in the investigations of human cognition.

Nevertheless, in spite of the necessity of higher level scientific inquire, the "hypothesized molecular mechanism is actually doing the causal work", i.e. "the best causal-mechanistic story for the specific cognitive function then resides at the lowest level of effective experimental interventions." (BICKLE, 2012, p. 104). This means that there is no explanatory pluralism and autonomy here, and this is in sharp contrast with the neo-mechanistic pluralist framework. For Bickle, what counts in the end for delivering the ultimate complete scientific explanation is the

molecular and cellular level. Not all levels are equally significant for the explanation, there is no pluralistic integration, but indeed a reductive one.

3 The mechanists' reply

The neo-mechanistic framework is committed to pluralism: its proponents argue for multilevel causal and explanatory integration in cognitive and neural sciences. Therefore, on the one hand, the framework rejects certain kinds of reduction. However, on the other hand, the framework assumes a kind of 'reductionist' stance. Bechtel claims that "from the point of view of mental activity" his approach is reductionist, and he calls it "*mechanistic reduction*" (2009a, p. 13-14 – highlighted in the original). As he states: "Mechanistic explanation, in seeking to explain the behavior of a mechanism in terms of the operations of its parts, is committed to a form of reduction." (BECHTEL, 2008, p. 129). Mechanistic reduction, in his view, occupies a middle ground between forms of dualism and Bickle's strong reductionism (cf. BECHTEL, 2008, p. 130).

The central problem with strong neuro-cognitive explanatory reductionism (Bickle's reductionism), in Bechtel's view, is due to the fact that "whole systems exhibit behaviors that go beyond the behaviors of their parts" (BECHTEL, 2008, p. 129). Therefore, the strong neuro-cognitive reductionist position is being characterized as defending the thesis that 'individual component parts can explain alone the behavior of a given whole'.

Bechtel, contrary to the extreme reductionist position that he characterizes, wants to be some sort of weak neuro-cognitive explanatory reductionist, since theories related to whole systems are at a higher explanatory level than theories related to their parts, given that wholes are more than their parts. Accordingly, "individual lower-level components do not explain the overall performance of the mechanism. Only the mechanism as a whole is capable of generating the phenomenon" (BECHTEL, 2008, p. 146). Craver's ideas on this point are very similar. In his view, since "mechanisms can do things that individual parts cannot" and "mechanisms explain things that individual parts cannot", thus "higher levels of mechanisms are legitimately included in the explanations of contemporary neuroscience" (CRAVER, 2007, p. 216). As he points out: "Mechanisms require the organization of components in cooperative and inhibitory interactions that allow mechanisms to do things that the parts themselves cannot do" (CRAVER, 2007, p. 216).

In the human neuro-cognitive mechanism of memory consolidation, for example, there are many levels of mechanistic compositional organization, such as the level of connections between systems of neural networks, particular inter-cellular processes, intra-cellular processes, and molecular processes. In each of these compositional levels, there are different causal processes occuring always at the same level, but these causal processes are mediated by the compositional relations, affecting all the different levels of the mechanism (cf. CRAVER; BECHTEL, 2007). Only when all the component parts and their interactions at a lower-level are considered, is it possible to understand the causal activity of the whole

mechanism at the higher-level. The entire neuro-cognitive mechanism is organized in a specific manner, and there are more causal processes occurring at this higher-level (that together are responsible for the behavior of the entire mechanism) than the causal processes at a lower-level related to individual (or sets of) parts. Thus, there is some independent causal higher-level – consequently causal plurality and causal and explanatory weak autonomy. There is here, therefore, causal and explanatory pluralism and weak autonomy, not strong neuro-cognitive explanatory reduction.[7]

Bechtel and Craver intend to claim that identifying components, operations and organization is important to explain the phenomena produced by the whole mechanism. But, at the same time, "the behavior of the whole system must be studied at its own level with appropriate tools for that level", since this level has "a kind of independence" and the phenomena are "different from those studied at the level of the component parts" (BECHTEL, 2008, p. 129). Therefore, the first line of argument is the emphasis on the importance of the internal organization of a given whole, which, in their view, already undermines strong neuro-cognitive explanatory reduction.

Another issue concerning the internal organization of a given whole is related to predictability, and it arises when we consider wholes which are very complex systems. There are systems which are not so complex, as for example, a bike, or a car, or even an airplane. If one wants to understand the functions and behaviors of such systems, all one needs to do is to look to their component parts and subparts to understand how they causally interact with each other. This knowledge about all the components of the system can largely explain what one needs to know about the system, and one can predict much of its behavior in this way, as long as the appropriate knowledge of external conditions is also provided.[8]

In highly dynamical complex systems, there is no linearity in the causal interactions and the components of the system relate not in a static, but in a dynamical way, which makes the system to change constantly its own state and the way it is related to the environment. In such systems accuracy in predictions is not so high. The complex interactions produce new phenomena that are radically different from those related to the components of the system and which cannot be fully predicted or explained just by the knowledge of the operations associated with those components taken in isolation. This occurs because the variables and their interactions are too many and they cannot be measured or understood fully. There will always be a degree of imprecision, a degree of uncertainty in the final value. And since the initial conditions of a system cannot be measured with complete accuracy, i.e. the initial measurement will always have a degree of uncertainty, the results derived from that measurement will also be uncertain.

[7] Craver appears to reject any form of neuro-cognitive explanatory reductionism (2007, p. 228ff.).
[8] To illustrate this, we can think about an airplane under turbulence, or a car being driven in a wet road on a raining day. In such cases, explanations concerning how these simpler systems will behave are often accurate, as well as the related predictions. Evidently, I am not claiming here that there are complete explanatory and one-hundred-percent accurate predictive models for simple systems in cases where the external relevant variables are too many.

As regards neural activity and cognition, if the brain (or parts of the brain) could be considered as such complex dynamical systems (BECHTEL, 2008, 2010), it would be impossible to predict cognitive phenomena arising from them just by looking at physicochemical neural operations and interactions of the components of this brain, or brain region. This is because it would be impossible to know all the initial conditions and to measure all the many variables in the brain responsible for the generation of a certain mental phenomena with total accuracy; it would be too complex. Consequently, any completely accurate prediction of complex human behavior related to mental phenomena would turn out to be impossible. As Craver points out:

> Some mechanisms have so many parts and such reticulate organization that our limited cognitive and computational powers prevent us from making [...] predictions. Some mechanisms are so sensitive to undetectable variations in input or background conditions that their behavior is unpredictable in practice. Behaviors of mechanisms are sometimes emergent in this epistemic sense. (CRAVER, 2007, p. 216-217)

This is another reason why strong neuro-cognitive explanatory reduction fails: certain mechanisms/systems/wholes with highly complex and dynamical internal organization behave in a way that cannot be predicted with high accuracy.

Moreover, as Bechtel emphasizes: "Both the level of the parts and the level of the mechanism engaging its environment play roles in mechanistic analyses. A mechanistic explanation therefore inherently spans at least these two levels." (2008, p. 148). This means that a particular whole mechanism behaves in a certain way "only under appropriate conditions." (BECHTEL, 2008, p. 146). Therefore, the context in which the mechanism behaves is another important aspect for understanding its behavior, not just the internal components, operations and their organization (cf. BECHTEL; ABRAHAMSEN, 2005, p. 426). It is thus crucial for a correct understanding of the behavior of a mechanism to identify the external factors that can affect it:

> No matter how much they investigate the parts, their operations, and their organization, investigators will not identify the variables in the environment that are impinging on the mechanism. Discovering these variables and their effects requires inquiry directed at the environmental variables using appropriate investigatory techniques (BECHTEL, 2008, p. 152).

This factors can be in turn understood as components in a larger mechanism, in which the target mechanism is embedded. It is this mechanistic environment that provides conditions for the rise of the particular behavior. Often, thus, an explanation must clarify what the appropriate environmental conditions for the appearance of a given phenomenon are.

Finally, it is also argued by neo-mechanists in cognitive science that so far there are no clear cases of strong neuro-cognitive explanatory reduction in the field. In their view, even Bickle's most compelling example related to the memory

system does not consist of a top-down search for lower-level mechanisms of memory – Bickle's picture, according to multilevel mechanists, is simply inaccurate and misleading. The example of memory can rather be understood in accordance with the pluralistic view of cognitive neuroscience and cognitive science integration (CRAVER, 2007, p. 237ff). In this view, the LTP neuro-physiological process is considered to be part of the explanation of memory consolidation, not as identical to memory consolidation, nor as a kind of memory consolidation. Accordingly, what Bickle is trying to do is to explain a whole based on its tiny parts. In other words, the particular molecular mechanism that involves cAMP, PKA and CREB, which leads to LTP is, thus, a tiny component of a larger memory mechanism for memory consolidation and cannot alone be considered the ultimate causal explanation for the phenomenon, as Bickle argues. Consequently, there are many important higher-level processes being left out in Bickle's reductive explanation.

Ultimately, therefore, the neo-mechanistic framework for cognitive science stands for pluralistic neuro-cognitive levels of causation and explanation, not for strong explanatory neuro-cognitive reduction.

4 The reductionists' rejoinder

Influential advocates of the neo-mechanistic framework for cognitive science such as Bechtel and Craver argue, thus, for quite basic and straightforward ideas: 1) that the behavior of a whole biological complex mechanism is more than the behavior of its parts taken in isolation; 2) that the organization involving all the components of a mechanism needs to be considered, if one wants to understand the behavior of the entire system; 3) that neuro-cognitive complex mechanisms can be unpredictable to some degree; and 4) that environmental external relations/factors influence the behavior of the entire system. All these points are presented as a case against strong neuro-cognitive explanatory reduction.

The strategy of the advocates of the mechanistic framework is to argue that since the system is organized in a specific manner, and this organization is a high-level aspect of the whole mechanism, because it refers to causal processes that are not just related to individual or sets of parts, but rather to the whole system, then there is some causally independent higher level – consequently, causal and explanatory plurality and weak autonomy.

Indeed, frequently in nature the operation of just one component, or a set of components smaller than the whole mechanism/system, cannot account for the behavior of the whole. For example, the operations of the motor of a car cannot account alone for the whole motion of the car. A whole often has features that none of its constitutive component parts have. This is trivially true, though. Similarly, a collection of *all the component parts* of a whole without considering its organization is also not enough to explain the whole behavior of the mechanism. If one takes just all the component parts of a car without considering how they are related within the car, it is impossible to explain how the car works. In this case, one would be talking about an 'aggregate', not about a whole, since

a mechanistic whole needs to have some organization to be so considered. Therefore, organization including all the components needs to enter the picture.

But Bickle does not argue that a single component's function, or a set of components' functions, always explain the function of a given whole mechanism. This would be a naive view, and Bickle does not take this position. A reductionist position of this kind has never been seriously defended in the specialized literature, because it is clear that the position is trivially false (cf. STEPHAN, 1992, p. 32). The issue here is that what Bechtel and others see as higher-level, for Bickle, can be simply described in terms of lower levels.[9] Bickle knows that organization is important, but in his view all the information about the organization can be also described in terms of lower levels, i.e. at the level of molecules and cells in the case of brain and cognition, where ultimately lies the most important causal mechanistic explanation for him (cf. BICKLE, 2012, p. 104). Theurer and Bickle state that quite directly and explicitly: "[...] as we descend downward through nested mechanisms, opening black boxes at progressively lower-levels along the way, we are uncovering mechanisms that are progressively more explanatory." (2013, p. 106). For Bickle, the organization of all the relevant component parts and their causal interactions can be described according to lower-level terms. The whole mechanism, in his view, is constituted by molecular interactions between a small number of cells, as in the case of the mechanism for LTP. Memory consolidation is a process performed and thus explained by this small molecular mechanism. Therefore, Bickle does not want to take out organization in any sense. This means that the sum of all the component's operations plus their organization in order to provide an explanation for the behavior of the whole mechanism can also accomplish a reduction, as long as all this is described in lower-level science.

Fazekas and Kertész (2011) point out that the possibility of describing all the information at lower levels follows from the very assumptions of the neo-mechanistic framework, because the whole mechanism is just the same as its component parts organized in a particular way. They claim that the mechanistic framework is not able to reduce higher levels to lower levels while maintaining the explanatory weak autonomy of the higher levels simultaneously. In their view, the mechanistic relationship of constitution/composition is not the *asymmetrical* 'part-whole' relation, but the *symmetrical* relation of 'parts + organization = whole', which is identity (FAZEKAS; KERTÉSZ, 2011, p. 373). For them, the ultimate goal of the mechanistic approach "is to explain how a system performs certain tasks by understanding how its parts organized in the right way perform the *very same* task." (2011, p. 372 – highlights in the original). Bechtel and Craver characterize 'lower-level' in terms of individual parts and, then, construct an argument in order to claim that organization is missing at the lower level. This argument against explanatory reduction based on the putative higher-level organization is, however, misleading, because:

> […] it attacks a straw man: it is true (though trivially) if one restricts the characterization of the lower level to the characterization of

[9] Bickle accepts the account of levels of mechanistic composition/parthood as described by prominent neo-mechanists (cf. THEURER; BICKLE, 2013, p. 105).

> individual entities and their behavior. However, we see no reason why the targets of this argument, reductive accounts, should restrict themselves to this 'light' way of characterizing lower-level (FAZEKAS; KERTÉSZ, 2011, p. 377).

There is, therefore, no causal higher levels here, since the same organization can be fully described in lower-level terms. Given that 'parts plus organization' is identical to the mechanism as a whole, the relevant causal explanation for the given phenomenon under investigation is just the same at the lower level, presented with a different vocabulary.[10]

In a similar argumentative line, Soom states that these different mechanistic levels are just "different levels of description" (2012, p. 655), but one can be fully reduced to the other since there is no causal process over and above the organized components of the given mechanism. At each level of description different vocabularies are used by different scientific fields approaching the phenomena at the respective level. But this does not mean that there are novel causal processes at each of these levels. As Rosenberg (2015, § 1) also points out: the neo-mechanists claim that there are higher-levels of explanation because it is possible to identify higher-level autonomous causal processes; at the same time, however, they demand a complete explanation in terms of the organized constitutive/compositional lower-level parts of the mechanism. This is clearly an inconsistency.

Furthermore, Theurer argues that compositional mechanistic relations are transitive, i.e. the information required for the explanation at one level is passed entirely to the other level, since the composition relation of the parts plus organization accounts exhaustively for the behavior of a given mechanism (2013, p. 306). In other words, if a mechanism, M_1, is reduced to a set of sub-mechanisms, M_2, and this set of sub-mechanisms, M_2, is in turn reduced to a set of sub-sub-mechanisms, M_3, then M_1 is directly reduced to M_3. And this occurs in mechanistic explanations because of the ontological and explanatory commitments of the framework: in order to explain the operation of a given mechanism, M_1, one needs to take into consideration all its parts plus organization, i.e. M_2, because M_1 is exhaustively constituted by M_2, as M_2 is exhaustively constituted by M_3. Given this transitive character of mechanistic explanations, lower-levels of molecular and cellular neural mechanisms necessarily explain directly higher-levels of neural

[10] The account of levels presented by neo-mechanists is very different from the account of levels presented by logical empiricists (cf. OPPENHEIM; PUTNAM, 1958). For the latter, theoretical levels of scientific domains are integrated by a deduction of laws. Thus, it is a matter of epistemic deduction: logical, or mathematical. For neo-mechanists, however, levels are integrated by ontic composition of the relevant biological mechanisms. But since these compositional levels can be described in natural language using particular words and concepts, it is correct to say that they can be described with different vocabularies, as neo-mechanists themselves sometimes do. Thus, in biological mechanisms one can consider the dimension of ontic levels of composition as well as the dimension of epistemic levels of description and, consequently, different levels of vocabularies. Moreover, since there is a relationship of complete composition between the whole mechanism and its parts plus organization, the vocabulary related to lower-level parts taken together (not in any isolation whatsoever) will always be referring to the whole too, thereby accounting completely for it in more detail.

systems and cognitive functions, because higher-level explanations become unnecessary after the explanatory work has been done at the lowest-level, which is more detailed and accurate. This is strong explanatory neuro-cognitive reduction. And this is why Theurer and Bickle claim that: "'New' mechanism, by the nature of its key resources, may be far more reductionistic than some of its proponents notice (or admit)." (2013, p. 109 – emphasis in the original).

Even in stochastic cases concerning complex systems, where one cannot have information about all the components due to practical impossibilities leading to unpredictability, there is strong explanatory reduction. There is indeed no conflict between unpredictability in complex systems and Bickle's reductionism. For it is irrelevant if at a certain point in time a phenomenon or a set of phenomena occur that cannot be completely explained or predicted with great accuracy on the basis of the initial state of the system due to the inaccuracy of the measures at the start. What really matters for Bickle's strong neuroscientific explanatory reduction is whether it is possible to describe these phenomena in causal lower-level mechanistic physical language – and then to identify as many mechanistic variables as possible, put all the information together and ultimately produce a scientific explanation of the whole complex system's behavior at hand. In this reductive picture the relevant processes required for the final explanation can be acquired from the lower-level of the components' operations and their interactions; then this information can be put together in order to explain all the causal interactions at the higher-level, since they would be equivalent.

Another line of answer the neo-mechanists provide is to appeal to the external (e.g. environmental, social, cultural) factors that arguably play a causal role at the higher-level of the whole mechanism in determining its behavior. As the argument goes, contextual information is not captured by just investigating the lower-level and its individual parts. In order to capture this information, the higher-level is necessary, and thus there is explanatory weak autonomy (not strong reductive integration) at the higher level. However, there are also counter-arguments for this view. Since the interactions at the higher level are also among mechanisms, i.e. the target mechanism as a whole and its environment, which is composed of other mechanisms, there is nothing that prevents all these whole mechanisms that interact to be further decomposed in sub-components and sub-functions. All of them can be thus described in lower-level vocabulary, as well as their relations to each other. In other words, the reduction to lower levels can be equally applied to mechanisms that interact with a target mechanism as a whole at a given higher level. Fazekas and Kertész (2011, p. 378) correctly argue that everything that is in the higher-level context of a whole target mechanism can be further decomposed and explained in lower levels. Since they are all mechanisms, the reduction should be applied to all of them, without exception.

The argument related to the role culture plays in the development and determination of human cognition is important (cf. BRUNER, 1990; LEITE, 2018, chap. 5), but within the framework of neo-mechanism it is untenable (cf. MILKOWSKI et al., 2018). It is an attempt to formulate, based on a kind of physicalism, a non-reductive explanatory account in similar lines to previous

problematic forms of emergentism (cf. KIM, 2006). All mechanisms are within a larger mechanism (cf. BECHTEL 2008, 2009c; CRAVER, 2007). For instance, the visual system is within the neural system, which is within the biological system, which is within an ecological system, and so on. In this way, all causal factors external to the target mechanism are causes from parts that belong to a larger mechanism, which in turn can also be decomposed and understood at the lower level. Moreover, since all mechanisms are completely composed by all the lower-level parts plus organization, and the internal causes are mediated by the compositional relation, the effects of such external causes will be felt at this lower level, and can be explained at this level if all the important parts are considered. For example, if particular areas of the brain affect the area V1 in the occipital lobe, these effects can be understood at the level of cells and molecules, as neuroscientific research has been constantly showing. The same kind of decomposition can be applied to these other particular areas, external to V1. In fact, the 'higher level' in mechanisms is merely an abstraction, if we are not considering just some part of the mechanism. In this case, there is no 'higher level'. Referring to a 'higher level', after considering all the lower level parts plus the organization, is simply misleading.

Finally, the neo-mechanists' arguments appealing to the fact that allegedly there are no real cases of neuro-cognitive reduction in the cognitive and neural sciences are not going to convince Bickle and other neo-reductionists. Bickle can simply maintain that his theory of strong neuro-cognitive explanatory reduction concerning the capacity of memory consolidation provides a plausible case where neuro-cognitive reduction occurs in the cognitive and neural sciences. He can also argue that this kind of reduction is not peripheral and that it is, on the contrary, ubiquitous in neuro-cognitive science. Therefore, those arguments won't succeed in order to undermine Bickle's position and save neo-mechanistic neuro-cognitive pluralist integration.[11]

5 The problem for neo-mechanists in cognitive science remains

This contemporary controversy between the neo-mechanists and the neo-reductionists has implications for establishing theoretical foundations and theoretical integration in the contemporary field of cognitive science (including here cognitive neuroscience). Therefore, it must be considered carefully by those interested in this scientific area.

The first important issue to consider in this regard is concerned with the *explanans*. The question here is whether the authors on these two sides have a common view on what successful scientific explanations in cognitive science are. Bechtel and Craver are leading neo-mechanists working in the field of

[11] It is important to note that, in this paper, I am pointing out theoretical problems. Empirical research is extremely useful to discuss them, but they will not solve these problems alone. Besides, there is already in the literature a substantial discussion of empirical research. Now it is important to discuss more deeply the theoretical problems. I do not think, therefore, that more empirical research will be useful to settle this issue, and thus it is of little help to add more.

neuroscience and cognitive science. They have been articulating in the last decades a view of scientific explanations for these and other fields of contemporary science (BECHTEL; ABRAHAMSEN, 2005; BECHTEL, 2008; MACHAMER; DARDEN; CRAVER, 2000; CRAVER, 2007). Bickle and other neo-reductionists are equally making claims about how explanations in neuroscience should be constructed (BICKLE, 2006; BICKLE; THEURER, 2013). While their accounts may differ on important aspects, all these authors agree with some basic notions. For instance, none of them subscribe to a view of scientific explanation based on logical derivation of laws and theories in a deductive argument – a model defended most prominently by some logical empiricists around the middle of the last century.

For these contemporary authors, properly understanding and explaining some natural biological phenomenon is a matter of providing a description of the biological mechanism responsible for generating this phenomenon. To explain here, thus, is to provide descriptions of working components and subcomponents of macro-mechanisms, their regular and irregular interactions and how the macro-mechanism is affected by external varying conditions. The macro-mechanism is considered to produce the phenomenon under investigation, i.e. what ultimately needs to be explained. The most accurate and detailed the account of the macro-mechanism and the external conditions that affect its functioning are, the better, because this will be more explanatory. Thus, there is a minimum common ground for comparing the explanatory frameworks in cognitive science theoretically, even if they present different views on some particular topics.

The second issue is related to the *explanandum* and to empirical evidence about its explanation. The question here is whether the authors in the debate are using examples of explanations in cognitive science concerned with the same phenomena, or very similar phenomena, that can be easily compared. This is indeed not the case. Bechtel (2009a) uses examples of episodic memory consolidation in humans, Craver (2002, 2007) uses the example of spatial memory consolidation in mice, and Bickle (2012) uses the example of memory consolidation for fear conditioning in mice. It would be certainly helpful if we could make these discussions and comparisons more systematic by using the same examples. In this way, one would be able to understand more precisely to what extent the different theoretical frameworks can be successfully applied in order to explain particular empirical phenomena, and which one presents a better explanation (cf. THEURER, 2013). Otherwise, while Bechtel and Craver will claim that the macro-mechanism amounts to *A*, *B* and *C*, Bickle will claim that the macro-mechanism amounts actually to *A*. In this case, the attempts of comparison will start by begging the question and reach nowhere. Nevertheless, apart from the difficulties in comparing empirical results to make a case for a particular framework, there is a more theoretical issue that remains a problem for the leading neo-mechanists.

Based on the discussions analyzed in the previous sections, one can note that the crucial point of disagreement is related to the relationship between levels in the mechanistic framework and which level is more explanatory. Bechtel and Craver (2007) claim that their account is 'causal at the intra-level' and 'constitutive

at the inter-level'. This is the so called 'mechanistic constitution' relation (CRAVER, 2007). In a recent publication, Povich and Craver state that "constitutive mechanistic explanations" are those in which "one can explain the behavior of the whole in terms of the organized behaviors of its parts, and one can explain the behaviors of the parts in terms of the organized behaviors of their parts." (2018, p. 186). They repeat that "the whole has capacities the parts alone do not possess" (2018, p. 190) and they say, furthermore, that: "If two things are not related as part to a whole, they are not at different levels" (2018, p. 188).

However, if one defines lower-level just in terms of what a part does and state, at the same time, that there can be no complete explanation of the behavior of a mechanistic whole based just on what parts do, then it is already *in principle wrong* to provide an explanation based on the lower level. The point articulated by the neo-reductionists is just that the lower level does not need to be defined in these terms. It is rather a matter of looking deeper into the sub-parts of the parts and provide an explanation based in that vocabulary, at that level of description, with all the required details. All the organization of the macro-mechanism, together with the behaviors of all the relevant parts, can have such description, at least in some cases.

Moreover, the mechanistic constitution idea appears indeed to lead to an identity relationship, as Fazekas and Kertész (2011) argue, since the mechanists are committed to the idea of explaining *completely and exhaustively* the function of a mechanism by its constitutive parts and organization in a given context. Povich and Craver state that: "The behavior of the whole contains the behaviors of the parts, and the behaviors of the parts collectively and exhaustively constitute the behavior of the whole." (2018, p. 193). If this is so, then there is no further space for anything at some higher level that cannot be described using a vocabulary at a lower level. Anything that is necessary for the explanation can be fully described:

> If one knows all of the relevant entities, activities, and organizational features, and knows all the relevant features of the mechanism's context of operation, and can in principle put it all together, then one must know how the mechanism will behave. […] It is an epistemic warning sign if features of the mechanism's behavior cannot be accounted for in terms of our understanding of its parts, activities, organization, and context. Mechanists thus operate with a background assumption that the phenomenon is exhaustively explained (in an ontic sense) by the organized activities of parts in context (POVICH; CRAVER, 2018, p. 193).

However, Povich and Craver claim that given the possibility of multiple realizability, token and type identity are not good terms to characterize their mechanistic relationship between levels (2018, p. 193). Another problem that they point out is that we do not have criteria to establish if a psychological function is indeed identical to another; thus, it is very hard to know if a psychological function is indeed identical to a function performed by a given mechanism. If it is so, then, this is as much a problem for 'identity relations' as it is for 'complete and exhaustive mechanistic constitution'. For if a psychological function, F_1, can be performed by

different macro-mechanisms, M_1, M_2 and M_3, then just providing a description of all the relevant activities related to M_1 will not be enough to provide an exhaustive description of the production of F_1. In case the number of macro-mechanisms responsible for F_1 is countable and their activities explainable, an identification between the psychological function and the number of mechanisms producing it is still possible. In case the number is uncountable, there can be no exhaustive constitution and thus explanation for that psychological function. The same happens if the psychological function cannot be identified as being the same function in different times considering at least the properties that are most relevant for that characterization. In this case, it is not possible to claim that a given macro-mechanism completely and exhaustively constitutes that psychological function, because there could be no certainty about that, or high probability that it is the case.

However, in case exhaustive mechanistic constitution is plausible, there is no good reason to suppose that the internal macro-organization of a whole human neuro-cognitive mechanism cannot be explained by lower-level components and their interactions, at least in some cases, such as memory consolidation as described by some neo-mechanists in neurophysiological terms. In other words, there is no reason to suppose that an account of the organization of the parts of a given macro-mechanism in some cases cannot be provided using the vocabulary of the lower level. The causal novelty or independence that is supposed to appear at the higher level is never clearly pointed out. But if it is present in the model, then one cannot claim that the behavior of the whole macro-mechanism will be entirely and exhaustively explained by all its component parts plus their organization in a particular context.

Someone defending the neo-mechanistic framework might still try to argue that the organization of a macro-mechanism can only be characterized with higher-level vocabularies, since the organizational feature of a mechanism is more than the spatial disposition of its parts: it includes complex causal interactions at different times between many operations and activities.

However, what I am showing in this paper is precisely that this argument is untenable. All the features related to the internal organization of biological mechanisms are compositional in nature. Consequently, as soon as they are completely decomposed by empirical research and all these decomposed parts and their operations are properly related in lower-level vocabulary, a complete account of the phenomenon under investigation will be provided at this lower level. There is no way out of this. Any attempt to show that there is a significant novelty at a supposed higher level will necessarily damage the mechanistic integration by exhaustive composition.

The source of the problem here is the emphasis from leading neo-mechanists in considering just parts in isolation when talking about lower levels. This creates the illusion that lower levels are always parts in isolation. But the lower levels of mechanistic composition are, on the contrary, all the parts and operations fully decomposed and taken together, i.e. organized together. Consequently, the organization is not just at a 'higher level'. It can be understood

and described at the lower level as well, because this organization (even if it is a very complex one) is simply the interactions between the parts (composed or fully decomposed) that exhaustively constitute the whole mechanism. Thus, even the most complex organization can be fully described in a lower-level vocabulary, as the research in the field of molecular and cellular neuroscience shows.

There is an internal contradiction within the neo-mechanistic framework for cognitive science: a full explanation of a given neuro-cognitive phenomenon requires all the mechanistic components plus the organization to be described, what can be completely and exhaustively done at a lower level, but if there is some novel and autonomous causal element at a higher level (given some sort of ontological emergence, or multiple realizability), no full explanation is possible at the lower level. This problem of the *inconsistence and instability in neo-mechanistic integrative explanations* is the central problem that remains as an obstacle for the pluralist integration defended by many supporters of the neo-mechanistic framework in cognitive science.

Concluding remarks

The neo-mechanistic framework has been very influential in cognitive science, as well as very significant for many important issues intensively debated in the field, such as those related to scientific integration.

However, the most central arguments the advocates of the framework use against strong neuro-cognitive explanatory reduction present significant shortcomings. Neo-mechanists argue that the strong neuro-cognitive reductionist is committed to the view that 'a part of a whole must explain the behavior of the entire whole'. That is not correct, however. Bickle is aware that individual parts do not explain the behavior of the whole in the life sciences. Instead, what he claims is that all the aspects of such explanation can be described at a lower level. This includes the organization of the whole biological neuro-cognitive mechanism, biological neuro-cognitive mechanisms with high degrees of complexity, and the external factors affecting the whole biological neuro-cognitive mechanism.

I am not presenting a new defense of neo-reductionism here. Bickle's framework is not being advocated here as the most plausible account for investigating human cognition in cognitive science. The strong neuro-cognitive reductionist framework presents indeed many problems (cf. e.g. LOOREN DE JONG; SCHOUTEN, 2005; LOOREN DE JONG, 2006). But it was not my purpose to analyse the plausibility of this neo-reductionist framework as a whole in this paper.

I am simply taking the view that neo-mechanists' defense of 'weak reductionism'/ 'weak pluralist integration' is inconsistent. The neo-mechanistic framework for the cognitive and neural sciences gives full space for Bickle to advance his approach. Underlying the neo-mechanistic framework is a strong neuro-cognitive reductionist program. This can be seen through the examination of the ultimate consequences of the theoretical commitments presented by

influential neo-mechanists. And Bickle is, evidently, taking full advantage of this (cf. THEURER; BICKLE, 2013).

The analysis shows, therefore, that, although the neo-mechanistic framework for cognitive science is allegedly committed to neuro-cognitive causal and explanatory pluralism, it is not able yet to provide a consistent defense of it, collapsing into a strong neuro-cognitive explanatory reductionist framework, or being an inconsistent framework. As a result, the neo-mechanistic framework's defense of neuro-cognitive explanatory pluralist integration is untenable. No robust, meaningful, or even weak explanatory autonomy of cognitive science can be achieved in this way.

References

BECHTEL, William; ABRAHAMSEN, Adele. Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, v. 36, p. 421-441, 2005.

BECHTEL, William; RICHARDSON, Robert. *Discovering complexity*: decomposition and localization as strategies in scientific research. Cambridge, MA: MIT Press, 2010. (Originally published in 1993).

BECHTEL, William; WRIGHT, Cory. What is psychological explanation? In: SYMONS, John; CALVO, Paco (eds.). *The Routledge Companion to Philosophy of Psychology*. New York: Routledge/Taylor & Francis Group, 2009. p. 113-130.

BECHTEL, William. Levels of description and explanation in cognitive science. *Minds and Machines*, v. 4, p. 1-25, 1994.

_____. Reducing psychology while maintaining its autonomy via mechanistic explanations. In: SCHOUTEN, Maurice; LOOREN DE JONG, Huib (eds.). *The matter of the mind*: philosophical essays on psychology, neuroscience, and reduction. Malden, MA: Blackwell Publishing, 2007. p. 172-198.

_____. *Mental mechanisms:* philosophical perspectives on cognitive neurosciences. New York: Routledge, 2008.

_____. Molecules, systems, and behavior: Another view of memory consolidation. In: Bickle, John (ed.). *The Oxford Handbook of Philosophy and Neuroscience*. New York: Oxford University Press, 2009a. p. 13-40.

_____. Constructing a philosophy of science of cognitive science. *Topics in Cognitive Science*, v. 1, n. 3, p. 548-569, 2009b.

_____. Looking down, around and up: mechanistic explanations in psychology. *Philosophical Psychology*, v. 22, n. 5, 543-564, 2009c.

_____. How can philosophy be a true cognitive science discipline? *Topics in Cognitive Science*, v. 2, p. 357-366, 2010.

_____. Explicating top-down causation using networks and dynamics. *Philosophy of Science*, v. 84, p. 253-274, 2017.

BICKLE, John. *Philosophy and neuroscience*: a ruthlessly reductive account. Norwell, MA: Kluwer Academic Press, 2003.

_____. Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, v. 151, p. 411-434, 2006.

_____. Real reduction in real neuroscience: metascience, not philosophy of science (and Certainly Not Metaphysics!). In: HOHWY, Jakob; KALLESTRUP, Jesper (eds.). *Being reduced*: new essays on reduction, explanation, and causation. Oxford: Oxford University Press, 2008. p. 34-51.

_____. A brief history of neuroscience's actual influences on mind–brain reductionism. In: GOZZANO, Simone; HILL, Christopher S. (eds.). *New perspectives on type identity*: the mental and the physical. Cambridge: Cambridge University Press, 2012. p. 88-110.

_____. Marr and reductionism. *Topics in Cognitive Science*, v. 7, p. 299-311, 2015.

_____. Revolutions in neuroscience: tool development. *Frontiers in Systems Neuroscience*, 2016. doi: 10.3389/fnsys.2016.00024.

BOONE, Worth; PICCININI, Gualtiero. The cognitive neuroscience revolution. *Synthese*, v. 193, n. 3, p. 1509-1534, 2016. doi:10.1007/s11229-015-0783-4.

BRUNER, Jerome. *Acts of meaning*. Cambridge, MA: Harvard University Press, 1990.

CRAVER, Carl F.; BECHTEL, William. Top-down causation without top-down causes. *Biology and Philosophy*, v. 22, p. 547-563, 2007.

CRAVER, Carl F.; TABERY, James. Mechanisms in Science. In: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), 2015. Available at: https://plato.stanford.edu/archives/spr2017/entries/science-mechanisms/.

CRAVER, Carl F. Interlevel experiments and multilevel mechanisms in the neuroscience of memory. *Philosophy of Science*, v. 69, n. S3, p. S93-S97, 2002.

_____. *Explaining the brain*: mechanisms and the mosaic unity of neuroscience. Oxford: Oxford University Press, 2007.

FAZEKAS, Peter; KERTÉSZ, Gergely. Causation at different levels: tracking the commitments of mechanistic explanations. *Biological Philosophy*, v. 26, p. 365-383, 2011.

GLENNAN, Stuart; ILLARI, Phyllis (eds.). *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge, 2018.

KAPLAN, David. Neural Computation, Multiple Realizability, and the Prospects for Mechanistic Explanation. In: KAPLAN, David (ed.). *Explanation and integration in mind and brain science*. Oxford: Oxford University Press, 2017. p. 164-189.

KIM, Jaegwon. Emergence: core ideas and issues. *Synthese*, v. 151, n. 3, p. 347-354, 2006.

LEITE, Diego A. *The twenty-first century mechanistic theory of human cognition:* a critical appraisal. Doctoral Dissertation. University of Trento (Italy), 2018. Available at: http://eprints-phd.biblio.unitn.it/2828/.

LOOREN DE JONG, Huib; SCHOUTEN, Maurice. Ruthless reductionism: a review essay of John Bickle's philosophy and neuroscience: A ruthlessly reductive account. *Philosophical Psychology*, v. 18, n. 4, p. 473-486, 2005.

LOOREN DE JONG, Huib. Explicating pluralism: where the mind to molecule pathway gets off the track – reply to Bickle. *Synthese*, v. 151, p. 435-443, 2006.

MACHAMER, Peter; DARDEN, Lindley; CRAVER, Carl F. Thinking about mechanisms. *Philosophy of Science*, v. 67, n. 1, p. 1-25, 2000.

MILKOWSKI, Marcin. Integrating cognitive (neuro)science using mechanisms. *Avant: Journal of Philosophical-Interdisciplinary Vanguard*, v. VI, n. 2, p. 45-67, 2016. doi:10.26913/70202016.0112.0003.

MILKOWSKI, Marcin et al. From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology*, v. 9, 2018. doi:10.3389/fpsyg.2018.02393

OPPENHEIM, Paul; PUTNAM, Hilary. The unity of science as a working hypothesis. In: FEIGL, Herbert; SCRIVEN, Michael; MAXWELL, Grover (eds.). *Concepts, theories, and the mind-body problem*. Minneapolis: Minnesota University Press, 1958. p. 3-36.

PICCININI, Gualtiero; BAHAR, Sonya. Neural computation and the computational theory of cognition. *Cognitive Science*, v. 37, n. 3, p. 453-488, 2013.

PICCININI, Gualtiero; CRAVER, Carl F. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, v. 183, n. 3, p. 283-311, 2011.

PICCININI, Gualtiero. Computing mechanisms. *Philosophy of Science*, v. 74, n. 4, p. 501–526, 2007.

_____. Computationalism. In: MARGOLIS, Eric; SAMUELS, Richard; STICH, Stephen P. (eds.). *The Oxford Handbook of Philosophy of Cognitive Science*. New York: Oxford University Press, 2012. p. 222-249.

_____. *Physical computation*: a mechanistic account. Oxford: Oxford University Press, 2015.

POVICH, Mark; CRAVER, Carl F. Mechanistic levels, reduction and emergence. In: GLENNAN, Stuart; ILLARI, Phyllis (eds.). *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge, 2018. p. 185-197.

ROSENBERG, Alex. Making mechanism interesting. *Synthese*, 2015. doi: 10.1007/s11229-015-0713-5.

SILVA, Alcino; BICKLE, John. The science of research and the search for molecular mechanisms of cognitive functions. In BICKLE, John (ed.). *The Oxford Handbook of Philosophy and Neuroscience*. New York: Oxford University Press, 2009. p. 91-126.

SILVA, Alcino; LANDRETH, Anthony; BICKLE, John. *Engineering the next revolution in neuroscience*: the new science of experiment planning. New York, NY: Oxford University Press, 2014.

SOOM, Patrice. Mechanisms, determination and the metaphysics of neuroscience. *Studies in History and Philosophy of Science*, v. C43, p. 655-664, 2012.

STEPHAN, Achim. Emergence – a systematic view on its historical facets. In: BECKERMANN, Ansgar; FLOHR, Hans; KIM, Jaegwon (eds.). *Emergence or reduction? Essays on the prospects of nonreductive physicalism*. Berlin: Walter de Gruyter, 1992. p. 25-48.

THAGARD, Paul. *Hot thought: mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press, 2006.

_____. Why cognitive science needs philosophy and vice versa. *Topics in Cognitive Science*, v. 1, p. 237-254, 2009.

_____. Cognitive Science. In: ZALTA, Edward N. (ed.). *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), 2018. Available at: https://plato.stanford.edu/archives/win2018/entries/cognitive-science/.

THEURER, Kari; BICKLE, John. What's old is new again: Kemeny-Oppenheim reduction in current molecular neuroscience. *Philosophia Scientiae*, v. 17, n. 2, p. 89-113, 2013.

THEURER, Kari. Compositional explanatory relations and mechanistic reduction. *Minds & Machines*, v. 23, p. 287-307, 2013.

VON ECKARDT, Barbara. *What is cognitive science?* Cambridge, MA: MIT Press, 1993.

WRIGHT, Cory; BECHTEL, William. Mechanisms and psychological explanation. In: THAGARD, Paul (ed.). *Philosophy of psychology and cognitive science*. Amsterdam: North Holland Elsevier, 2007. p. 31-79.

ZEDNIK, Carlos (2018). Mechanisms in cognitive science. In: GLENNAN, Stuart; ILLARI, Phyllis (eds.). *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge, 2018. p. 389-400.

Recebido em: 14-01-2019

Aceito para publicação em: 25-06-19

FISICALISMO E O PROBLEMA MENTE-CÉREBRO: UMA QUESTÃO DE DEFINIÇÃO

PHYSICALISM AND THE MIND-BRAIN PROBLEM: A MATTER OF DEFINITION

JULIO CÉSAR MARTINS MAZZONI[1]
Universidade Federal de Juiz de Fora (UFJF) – Brasil
julio.mazzoni@ich.ufjf.br

RESUMO: O Fisicalismo tem sido a posição filosófica monista mais aceita no *mainstream* do debate contemporâneo sobre a natureza do mental. Mas o que significa dizer que tudo o que existe é "físico"? O presente trabalho busca responder à pergunta: Como as teses fisicalistas contemporâneas têm definido o termo 'físico' em suas proposições? Para respondê-la foi realizada uma investigação teórico-filosófica baseada em revisão bibliográfica e análise lógica e conceitual. Quatro categorias gerais de definição do termo 'físico' foram identificadas numa revisão da literatura. Existem igualmente fortes críticas às propostas de definição em toda a discussão filosófica. Não temos uma definição incontroversa do que seja uma propriedade física, tampouco consenso sobre qual deveria ser a formulação mais adequada da tese fisicalista. Assim, a questão que se coloca é: Por que estamos discutindo o valor de verdade de uma tese que nem mesmo tem conseguido ser formulada de forma adequada?

PALAVRAS-CHAVE: Materialismo. Fisicalismo. Filosofia da Mente. Matéria. Físico.

ABSTRACT: *Physicalism is the most widely accepted position in the mainstream of contemporary debates about the nature of mind. However, what does it mean to say that everything is physical? This paper aims to answer the question: How do contemporary physicalists define the term 'physical' in their propositions? To answer this question I performed a bibliographic review together with conceptual and logical analysis. Four sorts of definition of "physical" were found, and all have been the target of strong criticism. We do not have an uncontroversial definition of what a physical property is, nor a consensus on what the most appropriate formulation of the physicalist thesis should be. So the question that arises is: Why are we discussing the truth-value of a thesis that has not even managed to be properly formulated?*

KEYWORDS: *Materialism. Physicalism. Philosophy of mind. Matter. Physical.*

> "Aquilo que se sabe quando ninguém nos interroga, mas que não se sabe mais quando devemos explicar é algo sobre o que se deve *refletir*"
> (WITTGENSTEIN, *Investigações Filosóficas*, § 89)

[1] Doutorado em Psicologia pela Universidade Federal de Juiz de Fora (UFJF).

Introdução

Qual é a natureza da mente humana? Seria a mente somente um produto derivado de processos cerebrais ou parte de uma alma imaterial interagindo com o corpo? Como o mental se relaciona com o físico? Tais questões intrigaram os seres humanos durante séculos e fazem parte hoje do chamado problema mente-corpo, discutido historicamente no âmbito filosófico e atualmente em áreas fronteiriças do conhecimento, como a Psicologia, a Neurociência e a Psiquiatra. É a questão ontológica central da Filosofia da mente, campo de investigação filosófica sobre a natureza do mental, e um dos tópicos mais debatidos na contemporaneidade.

O problema mente-corpo (atualmente mente-cérebro) já foi considerado um problema composto por um aglomerado de questões inter-relacionadas, contendo aspectos de natureza científica, epistemológica, sintática, semântica e pragmática, intimamente associados com questões filosóficas controversas como a teleologia, a intencionalidade e o livre arbítrio, assim como o extremo oposto, ou seja, um pseudoproblema (*Scheinproblem*), uma falsa questão (FEIGL, 1967).

Diferentes tentativas de respostas ao problema foram e continuam sendo oferecidas pelos mais diversos pensadores ao longo do tempo (ARMSTRONG, 1993; CHURCHLAND, 1981; DENNETT, 1996; FODOR, 1981; PLACE, 1956; POPPER & ECCLES, 1977; PUTNAM, 1967; RYLE, 1949; SEARLE, 1992; SMART, 1959; SWINBURNE, 1986). Qual é o estado do debate na atualidade? A possibilidade de uma resposta inequívoca para essa pergunta não é clara. Não obstante, temos um panorama, partindo de um ceticismo sobre a possibilidade de resposta ao problema, em função dos limites de nossa capacidade de conhecimento (McGINN, 1989), passando por inúmeras possibilidades de respostas concorrentes, até uma forte adesão ao Fisicalismo como melhor alternativa explicativa.

O fisicalismo tem sido a posição filosófica mais fortemente defendida e aceita no *mainstream* do debate atual sobre a natureza da mente (KIM, 2006), chegando a ser considerada a *Weltanschauung* da Filosofia contemporânea (GILLET & LOEWER, 2001). Numa recente pesquisa, Bourget e Chalmers (2014) constataram que 56% dos filósofos profissionais dos principais departamentos de filosofia do mundo consideram o fisicalismo a melhor opção frente à questão ontológica sobre o problema mente-corpo, 16% acreditam em outras alternativas e somente 27% responderam sustentar um anti-fisicalismo.

No estado atual do debate, os esforços têm focado na tentativa de demonstrar como a mente pode ser derivada de um estado físico, portanto, o escopo geral seria fornecer uma compatibilização do fenômeno mental com a visão de mundo fisicalista (MONTERO, 2001; KIM, 1998). Entretanto, igualmente, o fisicalismo tem sido alvo de críticas e questionamentos em razão das dificuldades teóricas encontradas para lidar com problemas filosóficos, tais como a consciência fenomenal e os *qualia*, a intencionalidade do mental, o problema do significado e a irredutibilidade do mental (CHALMERS, 1995, 1996, 2002; JACKSON, 1982, 1986; KOONS & BEALER, 2010; LEVINE, 1983; NAGEL, 1974; ROBINSON, 1996). O movimento de antítese às teses fisicalistas vem crescendo, não se restringindo

apenas à dimensão filosófica, mas também alcançando o domínio mais amplo da atividade empírica e científica (BEAUREGARD et al., 2014; GREYSON, 2010; PARNIA, 2007; SHELDRAKE, 2013; VAN LOMMEL, 2013).

Apesar de toda sua popularidade filosófica e relevância para o problema mente-cérebro, a pesquisa teórica/filosófica em âmbito nacional especificamente sobre o Fisicalismo é bastante incipiente, existindo somente um trabalho em português dedicado a revisar questões sobre a tese fisicalista (ZILIO, 2010) e nenhuma revisão conceitual sistemática sobre a definição do 'físico' nas teses fisicalistas. Portanto, o objetivo do presente trabalho é, a partir do recorte temático de um problema filosófico específico, a saber, a tese fisicalista e as definições do termo 'físico' pressupostas, realizar uma revisão das definições do 'físico' contidas nas teses fisicalistas dentro do debate contemporâneo em Filosofia da Mente, analisando suas implicações e consequências lógicas para o problema mente-cérebro. O nível de análise da pesquisa é puramente teórico-semântico, identificando a definição do 'físico' nas principais posições fisicalistas sustentadas atualmente no campo de discussão da Filosofia da Mente. Nossa metodologia consiste em uma investigação teórico-filosófica de clarificação com base em revisão bibliográfica e análise lógica e conceitual.

1 Fisicalismo: o materialismo ontológico contemporâneo

Mas, afinal, o que é o Fisicalismo? O termo 'fisicalismo' começa a ser utilizado na Filosofia na década de 1930 com as publicações dos positivistas lógicos Rudolf Carnap (1891-1970) e Otto Neurath (1882-1945) (CARNAP, 1959; NEURATH, 1931a, 1931b). A tese geral, ressalvando as particularidades em cada autor, é a de que qualquer proposição dotada de sentido é passível de ser traduzida, sem perda de significado, em afirmações utilizando uma linguagem física (daí o termo 'Fisicalismo'). É, portanto, uma tese linguística/metodológica (e não ontológica) relacionada a uma concepção unificada da ciência via critérios de verificação intersubjetivos da linguagem observacional (FEIGL, 1963), sem pretensões – ao menos explícitas – de fazer alegações de ordem metafísica. O Fisicalismo, na forma como concebido por Carnap e Neurath, porém, não permaneceu hegemônico. Autores como Quine (1954) e Smart (1963, 1981) contribuíram na propagação da ideia de que a tese fisicalista era de natureza metafísica e não linguística. Alguns têm deixado claro o fato do fisicalismo contemporâneo não ser mais entendido como uma doutrina metodológica, mas sim ontológica (PAPINEAU, 2001). Há resistentes ao uso intercambiável dos termos 'fisicalismo' e 'materialismo' para descrever concomitantemente a posição filosófica de que tudo o que existe é fisicamente constituído (BUNGE, 2010), embora muitos autores não utilizem mais o termo 'materialismo' nas discussões ontológicas, somente 'fisicalismo' (STOLJAR, 2010).

Independentemente de questões terminológicas e da discussão sobre sua associação com o materialismo, na forma como vem sendo compreendido atualmente, o termo 'fisicalismo' genericamente representa a proposição de que todas as coisas existentes são físicas, não existindo nada além do físico (STOLJAR,

2010). Considera, portanto, todos os processos ditos mentais submetidos a essa condição, seja por uma relação de identidade, seja por superveniência ou qualquer outro tipo de relação derivativa de bases fundamentalmente físicas. A partir da ideia fisicalista geral, diferentes distinções em suas modalidades e sentidos podem ser traçadas. Armstrong (1997) faz uma diferenciação entre duas versões, um fisicalismo forte e um fisicalismo fraco, tendo em vista a abrangência e a força das proposições. Nagel (1965) o divide em quatro tipos de formulações possíveis, desde uma afirmação de identidade específica entre condições mentais e sua contrapartida física, até versões mais brandas nas quais a identidade entre o físico e o mental é completamente flexibilizada. Fisicalistas contemporâneos reconhecem a possibilidade de expressão da "intuição geral" em múltiplos formatos, através da noção de redução, eliminação, superveniência física, realização física, identidade e assim por diante.

À primeira vista poderíamos pensar ser a popularidade do fisicalismo e sua ampla adesão consequência de um consenso oriundo de uma adequada formulação filosófica da tese. Entretanto, analisando mais de perto fatalmente somos confrontados com um panorama bastante diverso. Na realidade, pretendemos demonstrar ser mais aparente do que real o consenso em torno da proposição geral da tese fisicalista. Os filósofos têm apresentado diferentes definições gerais do que seria o fisicalismo (DOWELL, 2006b; FRANCESCOTTI, 2014; HELLMAN & THOMPSON, 1977; JACKSON, 1998; KIM, 2005; PAPINEAU, 2001; POLAND, 1994; SCHIFFER, 1990; STRAWSON, 2006). Se o consenso tem girado em torno da ideia intuitiva de que fisicalismo é a tese responsável por alegar não existir nada para além do físico, por outro lado, dificuldades e discordâncias surgem com relação ao próximo passo, ou seja, como especificar cada componente presente nessa proposição geral. Pode-se dizer que os elementos comuns entre as formulações são (a) a proposição afirmativa generalizante ("Tudo é", "A totalidade", "Todos os fatos", "Todas as verdades", "Não há nada além") ancorada na (b) noção de físico ("Exaustivamente compostas por propriedades físicas"; "Fundamentalmente físico"; "É realizada por entidades físicas"; "Em última análise compostas por propriedades físicas"). Em contrapartida, são grandes as divergências entre as teses quanto ao que se quer dizer exatamente ao usar o operador lógico 'Tudo'/ 'Para todo' e, especialmente, o que o termo 'físico' designa.

2 As distintas formas de definir e conceber o 'físico'

Conforme brevemente descrito anteriormente, segundo o materialismo ontológico/fisicalismo ontológico, todas as classes dos fenômenos existentes seriam constituídas fisicamente, não existindo nada além. Mas, o que significaria dizer que tudo o que existe é "físico"? Como o termo 'físico' presente nas teses fisicalistas deve ser compreendido? Essa pergunta, conhecida como a *Questão da Especificação* (CROOK & GILLET, 2001), tem rendido intensos debates nos últimos anos em um nicho restrito especializado em Filosofia.

Diante do problema da especificação do "físico" podemos, de um modo geral, identificar quatro categorias abrangentes nas tentativas de resposta à

pergunta: (I) Definições de apelo à Física presente, também conhecidas por *presentistas*, que alegam ser físico somente o que a Física atual postula em seu sistema teórico como tal (HELLMAN & THOMPSON, 1975, 1977; JACKSON, 1998, 2006; KIM, 2005; MELNYK, 1997, 2003; POLAND, 1994; SMART, 1978, 1981); (II) Definições de apelo à uma Física ideal, conhecidas por *futuristas,* cuja afirmação é de que físico é tudo aquilo contido numa compreensão completa do Universo de uma Física ideal do futuro (ARMSTRONG, 1997; DOWELL, 2006a; LOEWER, 2007; PETTIT, 1993); (III) Definição negativa, também conhecida como via negativa, a qual diz ser físico tudo aquilo que é não fundamentalmente mental (MONTERO & PAPINEAU, 2005; SPURRETT & PAPINEAU, 1999; WORLEY, 2006); e por fim, (IV) Definições híbridas e alternativas, composta pelos critérios de definição das opções anteriores combinados – por exemplo, uma definição futurista acrescida das restrições das definições negativas – e/ou restrições particulares para o que seria uma propriedade física (FEIGL, 1967; KIRK, 1996; MEEHL & SELLARS, 1956; NIMTZ & SCHUTTE, 2003; STOLJAR, 2001a, 2001b; VICENTE, 2011; WILSON, 2006; WITMER, 2016).

As definições de apelo à Física presente, ou definições presentistas do físico, geralmente identificam como físico todo o conjunto de entidades e propriedades fundamentais postuladas pela Física atual e/ou pela microfísica de partículas contemporânea. Portanto, estipular o que é o físico ou uma propriedade física nesta categoria de definição, ficaria a cargo da Física atual em seu sistema teórico-empírico. Podemos considerar autores como J. J. C. Smart, Hellman e Thompson, Jeffrey Poland, Andrew Melnyk, Frank Jackson e Jaegwon Kim como representantes dessa categoria conceitual.

J. J. C. Smart, forte representante do materialismo australiano, identifica a si mesmo como uma fisicalista, no sentido ontológico. Smart tornou-se famoso na Filosofia da mente por, em conjunto com U. T. Place, ter proposto, na década de 1950, a teoria da identidade mente e cérebro, alegando que cada estado mental particular é idêntico a um estado cerebral particular (SMART, 1959). Sua proposta consistia em defender o materialismo como uma teoria cuja reivindicação era de que não existe nada no mundo para além das entidades postuladas pela Física. Porém, em um primeiro momento, Smart fez uma observação disjuntiva relevante: o conteúdo fundamental da realidade, além do qual não existiria nada, poderia tanto ser as entidades determinadas como físicas pela Física atual, quanto as entidades que ainda viriam a ser postuladas no futuro por teorias físicas mais elaboradas (SMART, 1963, p. 651).

Em trabalhos posteriores, Smart (1978), buscando fornecer um conteúdo não disjuntivo e mais específico para seu fisicalismo, o definiu tendo por base somente a Física presente e atual, sem qualquer referência a entidades potencialmente postuladas por teorias físicas mais adequadas do futuro. O motivo oferecido pelo autor é que assim seria possível delimitar melhor o conteúdo do fisicalismo e, além disso, para os propósitos da Filosofia da biologia e da mente, seria suficiente e perfeitamente possível ancorar o fisicalismo aos princípios da Física presente. Posteriormente, Smart (1981) argumentou que sua posição seria um fisicalismo "não-definicional" e "não-dedutivo" (non-definitional and non-

deductive physicalism). Novamente em sua última proposta, Smart reitera que ao definir o conteúdo de seu fisicalismo com relação ao problema mente-cérebro, identifica por 'físico' os conteúdos da Física atual e suas teorias.

Outro representante dessa estratégia de definição (com apelo à Física atual) é Frank Jackson. A forma como define as propriedades físicas vincula-se à capacidade da ciência, especialmente da Física, em conseguir designar um conteúdo para o termo 'físico'. Para clarificar o conceito central da tese, ou seja, delimitar o que seria uma propriedade física, o autor identifica três características distintivas: (a) não ser senciente; (b) ser amplamente similar ("broadly akin") àquelas postuladas pela Física atual; e (c) ser necessária para lidarmos com e explicarmos o nível micro da realidade (JACKSON, 1998, p. 08). Mais recentemente, Jackson (2006) reforçou a ideia de um fisicalismo *a priori*, insistindo na possibilidade de definir o físico, para os fins fisicalistas, de uma forma abrangente ancorado, enquanto tese metafísica, na Física e nas ciências físicas em geral, tais como a físico-química, a bioquímica. Nesse sentido, Jackson consiste em outro proponente da estratégia de definição através da ciência física contemporânea como forma de delimitar o físico do fisicalismo.

Por outro lado, as definições de apelo à Física futura ou definições futuristas são todas aquelas que, diferentemente da abordagem anterior, buscam identificar o físico tendo por base não as ciências físicas do presente, mas uma Física ideal do futuro, completa e verdadeira, capaz de fornecer uma teoria final da totalidade das coisas existentes no universo. De modo geral, tais definições são acompanhadas de um reconhecimento de que a Física atual ainda está incompleta, imprecisa, sendo necessário apelar para outra via no processo de especificação do 'físico'. O que seria exatamente tal Física ideal do futuro, completa e verdadeira? Embora não seja uma questão autoevidente, sendo explicitada somente por alguns, de modo geral, o que a maior parte dos representantes dessa abordagem considera enquanto tal é uma teoria epistemologicamente ideal e verdadeira, num cenário futuro, dentro dos limites máximos de aquisição do conhecimento, levando em conta todos os dados empíricos possíveis sobre os fundamentos últimos da realidade. Exemplos de autores que adotaram essa estratégia são Philip Pettit, David Armstrong, Janice Dowell e Barry Loewer.

A título de exemplo, vejamos a versão do fisicalismo, chamada de Microfisicalismo, apresentada por Pettit (1993). O autor considera central o compromisso com duas proposições gerais básicas: 1) O mundo empírico é constituído por entes que a Física está em melhor condição de identificar; e 2) O mundo empírico é governado por forças e regularidades que a Física está melhor equipada para descrever. Pettit descreve ainda quatro alegações adicionais, compostas por afirmações particulares, coerentes com as suas duas proposições centrais:

(I) Existem entidades microfísicas: (a) Existe um mundo empírico no formato daquele postulado pela Física; (b) Diferentes tipos de coisas no mundo empírico compartilham uma composição em seu nível subatômico no formato daquele postulado pela Microfísica – um domínio de minúsculas entidades microfísicas.

(II) Entidades microfísicas constituem tudo o que existe: (a) Tudo no mundo empírico ou é composto, de algum modo, por entidades microfísicas postuladas pela Física ou não é composto, mas é um ente elementar em si; (b) Tal composição é conservadora, ou seja, duas entidades microfísicas não podem diferir intrinsecamente sem alguma diferença na configuração de seus componentes microfísicos.

(III) Existem regularidades microfísicas: (a) Entidades microfísicas estão submetidas a leis regulares em virtude de suas propriedades e relações microfísicas; (b) As leis operantes no domínio microfísico não são obtidas necessariamente por leis obtidas no nível macro. Podem ser leis iguais (conservação) ou diferentes, pois as leis microfísicas são primitivas.

(IV) As regularidades microfísicas governam tudo o que existe: (a) O nível macro não complementa as leis do nível micro; e (b) as leis macrofísicas não são independentes das leis do nível micro, ou seja, não têm o potencial de entrar em conflito com elas.

As quatro alegações supracitadas forneceriam suporte à ideia de que o mundo empírico contém somente o que uma Física verdadeira e completa afirmaria conter (PETTIT, 1993, p. 213). Apesar de sua formulação fisicalista deixar a cargo de uma Física completa designar o que é físico, assim como suas propriedades, Pettit (1993) estipula a existência de entidades microfísicas e determina restrições formais às relações de tais entidades com os demais domínios da realidade – uma Física completa e ideal descreveria o que são tais entidades microfísicas.

David Armstrong consiste noutro autor que podemos relacionar à abordagem que busca definir o físico através da ciência física ideal e completa. Armstrong, em seu *A World of states of affairs* (1997), defende que tudo o que existe no espaço-tempo são entidades físicas governadas por leis físicas. Ao considerar que os únicos particulares contidos no sistema espaço-temporal são entidades físicas e, mais precisamente, as leis físicas de uma Física completa, Armstrong adota a estratégia de postular o físico ancorando-o numa Física ideal e completa: "It asserts that the only particulars that the spacetime system contains are physical entities governed by nothing more than the laws of physics. The thesis is to be understood as a thesis about a completed physics" (ARMSTRONG, 1997, p. 06).

Armstrong, em sua delimitação da tese combinada com seu naturalismo e factualismo, considera o fisicalismo passível de ser compreendido como a tese que afirma ou 1 ou 2:

(1) Todos os universais fundamentais, sejam propriedades ou relações, são aqueles estudados pela Física, e todos os outros universais de primeira ordem são estruturas que implicam nada além dos universais fundamentais.

(2) Todas as leis fundamentais são conexões entre os universais fundamentais e outras leis, as quais não são nada além do que as leis fundamentais operando sob condições específicas e fronteiriças.

Embora apresente uma definição de natureza ontológica, o autor considera o fisicalismo um tipo de hipótese científica de alto nível, especulativa e reducionista. O fisicalismo, enquanto tese, extrairia sua autoridade da própria ciência – aqui restrita mais especificamente à Física.

Já no caso da definição negativa, também conhecida por via negativa, a proposta é definir o físico como tudo aquilo que é não mental, equacionando 'físico' com 'não mental' (MONTERO & PAPINEAU, 2005; SPURRETT & PAPINEAU, 1999). Alguns autores representativos dessa abordagem são David Papineau, David Spurrett, Barbara Montero e Sara Worley. Essa forma de definição é conhecida por via negativa justamente por buscar delimitar o físico através de uma definição do tipo negativa, ou seja, uma forma de definir algo por contraste com uma noção contrária.

David Papineau (1993; SPURRETT & PAPINEAU, 1999) é um dos maiores representantes da resposta pela via negativa quando o assunto é tentar delimitar o que seria o 'físico' nas teses fisicalistas. Um dos famosos argumentos a favor do fisicalismo, o argumento causal da completude física (Completeness of Physics), é considerado uma base crucial em sua defesa de uma definição do físico por contraste negativamente com o mental. Formalizando mais especificamente o argumento da completude da Física sustentado pelo autor a favor do fisicalismo, temos:

(P1) Todos os efeitos físicos são completamente determinados por leis e ocorrências físicas anteriores.

(P2) Se P1 é verdadeiro (a completude da física) e todos os efeitos físicos são devidos a causas físicas, então tudo o que tenha efeito físico deve em si mesmo ser físico.

Logo, não existe espaço para qualquer coisa não física fazer diferença, em termos de efeitos causais físicos.

Mas o que seriam tais efeitos e ocorrências físicas? Spurrett e Papineau (1999) deixam clara sua posição quanto à definição do 'físico' ao defender que nenhuma especificação detalhada, seja a partir de uma Física ideal, completa e verdadeira do futuro ou até mesmo da Física atual, consiste em condição necessária para caracterizar a proposta fisicalista. A tarefa principal ao formular o fisicalismo, no entendimento dos autores, é muito mais de excluir categorias especiais (tais como fenômenos mentais para além do físico) do que do incluir. Garantir que o termo 'físico' não designe qualquer tipo de propriedades e eventos não físicos ou sobrenaturais já seria suficiente para satisfazer a condição necessária expressa pelo fisicalismo.

De forma semelhante, não considerando ser possível definir adequadamente o físico, seja fazendo referência à Física presente, seja a uma Física ideal do futuro, Montero (2011) adota como rota alternativa a definição negativa. Para capturar a ideia associada à tese ontológica fisicalista, a autora afirma ser possível definir o que são propriedades físicas através da delimitação dos tipos de

propriedades que elas excluem. Isto é, as propriedades físicas seriam fundamentalmente não mentais, não divinas e não normativas, pois assim seria possível adequadamente fazer jus à inconsistência do fisicalismo com a existência de almas imateriais e propriedades mentais para além do domínio físico. O motivo apresentado, portanto, de uma adesão à via negativa seria o de que o ponto central de discordância entre fisicalistas e todos seus demais adversários no debate sobre a natureza da mente, em última instância, resume-se a questão da existência ou não de alguma característica especial e distintiva dos seres humanos no mundo natural (MONTERO, 2005).

Por fim, o que denominei de definições híbridas e alternativas consiste num conjunto mais recente de definições com características próprias, geralmente combinando aspectos das categorias gerais anteriores ou acrescentando elementos restritivos na especificação do físico não contidas nas categorizações anteriores. Desse modo, a categoria de definições híbridas e alternativas ou englobam definições que compartilham aspectos das categorias anteriores, porém, com novos elementos e/ou restrições, ou são totalmente distintas, idiossincráticas e alheias às abordagens anteriores. Autores passíveis de serem categorizados de tal forma são Meehl e Sellars, Herbert Feigl, Daniel Stoljar, Jessica Wilson, Robert Kirk, Agustín Vicente, Christian Nimtz e Michael Schütte, Witmer, dentre outros.

Um exemplo de tais definições híbridas do físico é a proposta de Jessica Wilson (2006). Denominada de NFM (no fundamental mentality), a qual, como o nome sugere, consiste em adicionar a restrição de negação da fundamentalidade do mental na definição do físico enquanto apela para a Física fundamental do futuro em sua definição. Sua proposta de definição pode ser formalizada da seguinte forma:

Def. Físico (No fundamental mentality, NFM): Uma entidade é física se, e somente se, satisfazer as seguintes condições:

I. É considerada como tal por uma Física fundamental do futuro; e

II. É fundamentalmente não mental.

O aspecto distintivo dessa estratégia é sua mescla de um apelo à Física futura com um critério restritivo do componente definicional central do fisicalismo. Combinando o apelo para a Física fundamental ideal como a ciência responsável por definir o que seja o 'físico' do fisicalismo, com tal restrição, a autora sustenta ser possível oferecer uma formalização mais adequada da tese. A relevância de adicionar a restrição de não fundamentalidade do mental visa evitar que o fisicalismo, pautando sua definição do físico numa Física do futuro, deixe em aberto a possibilidade para que essa mesma Física possa postular entidades e propriedades proto-fenomenais (mentais) enquanto fundamentais. Tal adição, segundo a autora, preservaria o fisicalismo enquanto tese antidualista de acordo com a tradição materialista, delimitaria seu conteúdo e garantiria a confiança geral

atribuída à Física como responsável por melhor identificar as entidades fundamentais da realidade.

De forma semelhante, Robert Kirk (1996) descreve por 'físico' o que quer que a Física postule enquanto tal, ao mesmo tempo em que busca excluir todas as expressões que podem ser consideradas psicológicas ou mentais. Recentemente, reforçando tal posição, e sem maiores delongas com relação à tarefa de definição do físico, Kirk (2013) afirma compreender o termo 'físico' de acordo com uma Física verdadeira e idealizada ('imagined true physics'), a qual não invoca fundamentalmente a consciência, intencionalidade ou qualquer outra noção psicológica na descrição e explicação das bases fundamentais do ser humano e demais organismos.

Nossa categorização ocorre não tendo como escopo esgotar todas as possibilidades de distinção e sistematização do conceito, mas somente de revisar panoramicamente como ele vem sendo discutido na Filosofia da Mente contemporânea, fornecendo assim, subsídios para uma análise mais acurada e derivações lógicas mais restritas. Em cada uma dessas categorias proliferam tentativas e propostas específicas de delimitação das condições necessárias para uma definição adequada do que venha a ser o "físico" presente nas teses fisicalistas. Apesar dos esforços, a questão da definição do termo 'físico' tem se evidenciado um problema filosófico de difícil solução.

Dentre as principais críticas ao problema da especificação encontramos a objeção da circularidade, o Dilema de Hempel, a crítica de Chomsky, de Bas Van Fraassen e de Chris Daly/ Israel Scheffler. A seguir iremos brevemente descrever algumas das críticas centrais aos esforços de delimitação do físico na formulação do fisicalismo contemporâneo.

3 Objeções e obstáculos à definição do 'físico'

Inicialmente, podemos pensar existir muitas opções, nessa miríade de definições, capazes de serem integradas à tese fisicalista e sua defesa da primazia do físico sobre o mental no debate mente-cérebro. Apesar de tais esforços, a pergunta "O que é o físico?" é bastante problemática e de difícil resposta.

O problema da circularidade (STOLJAR, 2010, 2016) consiste na imputação de circularidade na especificação das propriedades físicas em algumas definições, ou seja, ao tentar explicitar o que seria uma propriedade física invocando uma outra noção também considerada física (um objeto físico ou uma teoria sobre o que é físico), a caracterização do físico seria circular. A circularidade ocorre quando em uma definição de um termo X o conteúdo designado previamente já pressupõe X, ou seja, aquilo que busca definir. Em outras palavras, é um tipo de definição incapaz de esclarecer o que deseja definir, pois acaba por incluir na definição o termo a ser definido. Portanto, uma definição circular possui em seu *definiens* o próprio *definiendum*. Por exemplo, imaginemos uma proposta cujo objetivo seja definir o 'físico' através da seguinte definição:

Def. Físico: Tudo aquilo que em nosso discurso natural denominamos como objeto físico.

O exemplo, embora exageradamente grosseiro, ajuda a traduzir com mais clareza um tipo de definição circular, bem como tornar evidente o problema lógico enfrentado por esse tipo de definição. Em muitos casos, uma definição pode ser circular de forma mais tácita, quando apela para termos correlatos ou sinônimos daquele a ser definido. Portanto, a crítica de circularidade na definição do 'físico', nesse sentido, se aplicaria a qualquer definição que faz o uso, implícito ou explícito, de noções já pressupostas como físicas na definição do próprio físico. Esse tipo de definição não seria apropriado, à medida em que não delimita e estipula um predicado específico capaz de caracterizar aquilo que seria o conteúdo central da tese fisicalista.

Outra objeção, o Dilema de Hempel (HEMPEL, 1969; 1980), amplamente citado na discussão em torno do problema da especificação, seja como crítica, seja como obstáculo a ser superado, pode ser sintetizado da seguinte forma: Se o conceito de físico for baseado na Física contemporânea, o fisicalismo provavelmente é falso, porque as noções de físico têm sofrido historicamente significativas alterações e não existem razões para acreditarmos ser a Física atual uma descrição fiel e completa da realidade, muito pelo contrário. Por outro lado, se a noção de físico for ancorada numa Física ideal, uma Física futura completa, por consequência, seu conteúdo se torna obscuro e vago na medida em que seria impossível predizê-lo. Além disso, soma-se a essa última parte do dilema o que ficou conhecido como o *problema do pampsiquismo* (STOLJAR, 2016). Apenas apelar na definição de físico para propriedades e objetos paradigmaticamente físicos, não inclui nenhuma referência ontológica sobre os mesmos, assim, deixa em aberto a possibilidade de uma Física completa e ideal do futuro chegar à conclusão de que propriedades mentais fazem parte da natureza fundamental da realidade. Caso esse cenário hipotético aconteça, a Física estaria propondo a tese pampsiquista e/ou suas variações, o que levanta a questão se seria possível o fisicalismo na forma como é entendido hoje ser compatível ou não com a existência de propriedades mentais fundamentais e irredutíveis. Nesse caso, o problema é que tradicionalmente o pampsiquismo é uma tese antagônica ao materialismo ontológico, ou seja, definir o físico de modo que a proposição materialista seja compatível ou afirme a tese pampsiquista consistiria justamente numa tese não fisicalista, por definição, não sendo uma alternativa enquanto critério de definição capaz de delimitar as condições mínimas para uma tese fisicalita.

As observações de Hempel tiveram como resultado uma série de controvérsias quanto à adequabilidade da definição do físico e ainda hoje são motivo de acirradas disputas teóricas. Consiste numa crítica que obteve um amplo alcance, sendo reconhecida, seja por detratores, seja por defensores do fisicalismo (CRANE & MELLOR, 1990; HELLMAN, 1985; MONTERO, 2005; POLAND, 1994) como um problema sério a ser evitado ou superado nas definições do físico.

A crítica de Chomsky (1972, 1994) sinaliza dois aspectos problemáticos nas tentativas de formulação de um conceito do físico, a saber, a trivialidade da tese

fisicalista e a sua vagueza conceitual. O primeiro componente de sua crítica consistiria justamente, de certo modo, no problema lógico de a tese ser trivialmente verdadeira em função da forma com o físico é definido. Seu conteúdo seria maleavelmente adaptado aos avanços da atividade científica, sempre aberto a revisão, ou seja, qualquer reformulação da noção de físico postulada pela Física seria integrada, como já foi em outros momentos da história da ciência, ao *framework* fisicalista a partir de uma revisão do conceito, tornando o fisicalismo trivialmente verdadeiro.

Chomsky destaca, além do problema da trivialidade, o problema da vagueza, ou seja, o fato de conceitos como "físico" e "material", pressupostos em toda discussão sobre a natureza do mental, serem desprovidos de um sentido claro. Evidentemente, a vagueza em torno dos termos 'mental' e 'físico', por consequência, favoreceria justamente o materialismo usar o termo 'físico' para acomodar qualquer fenômeno, reforçando a possibilidade de a tese fisicalista ser trivialmente verdadeira. Afirmar, por exemplo, que tudo o que existe é físico, pressupondo que tudo o que a Física irá identificar quando estiver completa será físico, consistiria numa tese trivialmente verdadeira. Sua conclusão é de que na ausência de clareza quanto à definição e distinção do conteúdo de tais termos, aparentemente não seria possível a existência de qualquer proposta materialista, em seu sentido ontológico, e até mesmo do próprio problema mente-cérebro.

A crítica de Van Fraassen (1996) compartilha aspectos das anteriores e possui como eixos centrais: (a) a indeterminação do conteúdo da tese fisicalista; (b) a impossibilidade de falseamento; e (c) a defesa do argumento de que o fisicalismo não é uma hipótese científica, nem sequer uma tese filosófica, mas sim uma atitude. Fraassen considera terem fracassado as tentativas de resolver o problema da falta de conteúdo e clareza da tese, não conseguindo fornecer substancialidade ao fisicalismo. Segundo o próprio autor, o fisicalismo adotaria uma crença na veracidade de um "não sei o quê" baseada em um otimismo corajoso. Essa indeterminação do conteúdo, por seu turno, estaria diretamente relacionada à impossibilidade de falseamento da tese, pois uma vez que se compatibiliza sempre o significado de físico com os achados da Física mais atual, seja o que for por ela postulado, os novos dados empíricos nunca seriam capazes de contradizer a tese.

Segundo o autor, o materialismo seria uma tradição filosófica com uma roupagem diferente em cada era, com suas próprias reivindicações empíricas e teóricas, mantendo um núcleo de atitude e convicção invariante, ao qual Fraassen denomina de "o espírito do materialismo". Independentemente da tese, seu formato, termos e definições adotadas, o materialismo em cada época permaneceria sempre sobrevivendo, adaptando e se reinventando. Ou seja, embora os fisicalistas acreditem estar sustentando uma tese ontológica ou hipótese científica, na realidade, estariam expressando uma atitude sustentada em confusas condições desprovidas de maior clareza conceitual.

Em poucas palavras, não existiria, na forma como a tese vem sendo formulada, critérios adequados de verificação de sua falsidade. Existiria sempre a possibilidade de, mediante a ausência de evidências favoráveis, surgir uma nova versão da tese ou um enfraquecimento da mesma para acomodar os dados

empíricos produzidos pela atividade científica e, portanto, afirmações empíricas específicas não seriam capazes de desacreditá-la completamente. Por fim, Fraassen considera que o materialismo (entendido como sinônimo de fisicalismo ontológico) não pode ser considerado uma tese, mas sim um conjunto de atitudes, uma tradição filosófica, a qual em cada época da história tem desenvolvido a capacidade de ajustar seu conteúdo, recuar nas alegações e se remodelar como proposta filosófica. Esse seria o "espírito filosófico materialista", uma atitude filosófica e não uma tese.

As críticas de Chris Daly (1998) e Israel Scheffler (1950) caracterizam-se pela afirmação de que até a presente data não existe nenhuma distinção, baseada em princípios claros e bem definidos, entre o que seja uma propriedade física contrastante com as demais propriedades existentes, tendo com isso, duas implicações: (a) o fisicalismo não é uma tese bem definida, e (b) o debate entre o fisicalismo e dualismo, especificamente sobre o valor de verdade das teses, não faz sentido mediante a falta de clareza de um conceito fundamental presente na formulação de ambas.

Chris Daly (1998), ao investigar dentro dos domínios da metafísica das propriedades a seguinte questão "O que são propriedades físicas?", deriva uma conclusão bastante similar à de Scheffler por uma linha argumentativa diferente. Israel Scheffler, em seu trabalho de 1950, *The New dualism: Psychological and Physical terms*, analisa e faz críticas à distinção pressuposta entre as noções atribuídas aos termos 'psicológico' e 'físico', tendo por base os principais membros do círculo de Viena. A consideração de Scheffler é de que as tentativas dos positivistas lógicos, por ele analisados, de traçar uma linha divisória clara entre os termos 'físico' e 'psicológico' não foi capaz de obter sucesso. E de forma mais pungente, conclui não apenas que todas elas eram inadequadas, mas também que os esforços nessa direção eram inúteis. Já Daly, mais recentemente, buscou identificar quais são as bases de distinção – seus princípios, critérios de classificação, delimitação – pressupostas por algumas das principais abordagens ontológicas na diferenciação da classe de propriedades físicas com relação às demais propriedades existentes. Ao avaliar o status das principais abordagens – na atualidade – que visam estabelecer tais distinções (a saber, a proposta de Geoffrey Hellman**,** Poland, Papineau, Armstrong, bem como as definições de apelo à física atual e a uma Física ideal, completa e verdadeira do futuro) argumenta serem todas elas insatisfatórias. Ao analisar o problema de como distinguir a classe de propriedades físicas contingentes das demais, argumenta que tal fronteira não pode ser traçada, fazendo referência às propriedades existentes no mundo real, pois ainda que seja o caso de todas as propriedades do mundo real serem físicas, ainda seria necessário saber: (a) o que torna uma propriedade, uma propriedade física; e (b) dentre as propriedades possíveis, o que faz com que algumas delas sejam físicas e outras não. Sendo assim, o fisicalismo precisaria oferecer uma distinção, dentre todas as propriedades reais e possíveis existentes no universo, de quais seriam propriedades físicas, assim como sob os mesmos aspectos, quais não seriam.

Após sua minuciosa análise das principais abordagens existentes candidatas a estipular uma distinção entre uma propriedade física e as demais, sua conclusão,

portanto, é a de que não existe ainda nenhuma noção clara e bem definida do que seja uma propriedade física. Além disso, o autor levanta a questão se realmente existiria um princípio de distinção entre propriedades físicas e as demais existentes na natureza. A conclusão de Daly é imediata e pontual: não existiria nenhum princípio e critério de distinção bem definido entre o que seria uma propriedade física capaz de diferenciá-la de qualquer outra propriedade. Por consequência, considera intratável a questão da distinção do que seja uma propriedade física, chegando a afirmar, por fim, a necessidade de abandono dos programas metafísicos que pressupõe tal distinção. Como consequência, pressupor a existência de tal distinção é um equívoco, a noção de propriedade física não estaria bem definida inexistindo explicações satisfatórias do que ela seja até o presente momento. Tais problemas de ordem conceitual na formulação da noção de físico dissolveriam o problema mente-cérebro.

Além dos autores supracitados, outros têm reforçado as fileiras de críticas ao fisicalismo, enfatizando a ausência de uma divisão evidente entre o mental e não mental, a falta de uma noção clara do que seja um objeto/propriedade físico ou a trivialidade lógica em algumas versões do fisicalismo. Objeções e dificuldades com relação às tentativas específicas de definição e solução do problema também foram identificadas (CRANE & MELLOR, 1990; FRANCESCOTTI, 2014; SCHAFFER, 2003; STURGEON, 1998). Diferentes iniciativas de superação das críticas acima mencionadas e apresentação de uma definição final do físico têm sido evidenciadas na literatura, assim como movimentos de fortalecimento das críticas já existentes (BISHOP, 2006; BLUMSON & TANG, 2015; BOKULICH, 2011; DOWELL, 2006a; GILLETT & WITMER, 2001; HELLMAN & THOMPSON, 1975; JUDISCH, 2008; MELNYK, 1997, 2001, 2003; MONTERO, 2005; MONTERO & PAPINEAU, 2005; MSIMANG, 2015; NEY, 2008; PAPINEAU, 1991; POLAND, 2003; SMART, 1978; SPURRETT, 2001, 2015; STOLJAR, 2010, 2001a, 2001b). Contudo, o fato de ainda ser possível identificar nas últimas décadas diferentes produções buscando discutir e propor definições da tese fisicalista, bem como publicações recentes sugerindo uma "nova definição de físico" (WITMER, 2016) e do fisicalismo (DOVE, 2016), são evidências suficientes para demonstrar a ausência de consenso dos critérios definicionais, devido às discordâncias relativas àquilo que a proposição irá abarcar, restringir e denotar pelo termo 'físico'. Ademais, até a presente data não está claro o quanto e se tais objeções foram adequadamente respondidas e superadas.

4 Impasses e implicações derivadas do problema da definição do 'físico'

Tendo por base tudo o que foi apresentado até aqui, podemos estabelecer algumas considerações sobre o problema da definição do físico. Partindo de um predicado F – para o termo 'físico', temos:

(I) Se o predicado F faz referência a propriedades descritas pela Física contemporânea/atual, resulta em (a) predicações temporárias sujeitas a alterações, radicais ou parciais, com o desenvolvimento futuro da Física; e/ou (b) aumento da probabilidade da tese fisicalista ser falsa em sua formulação devido às limitações

da Física atual; e/ou (c) numa definição inapropriada, incapaz de abarcar a extensão factual do conceito por ela designado;

(II) Se o predicado F faz referência a propriedades descritas por uma Física ideal, completa e verdadeira no futuro, resulta em (a) vagueza conceitual; e/ou (b) trivialidade lógica da tese fisicalista; e/ou (c) possibilidade de, em última instância, dependendo do que uma teoria final da Física identificar, transformar-se num pampsiquismo ou ser identificada com demais teses rivais;

(III) Se o predicado F identifica o físico negativamente enquanto propriedade 'não mental' resulta em (a) não definir de forma clara e positiva o que é o mental; e/ou (b) na incapacidade de excluir teses rivais, devido a sua potencial compatibilidade com a noção de elán vital, o que não seria nem físico, nem mental, por exemplo (STOLJAR, 2016); e/ou (c) numa definição excessivamente restritiva com relação a algumas propriedades tidas como existentes enquanto 'não físicas' (numerais, valores, etc.); e/ou (d) num deslocamento de alguns obstáculos da definição do 'físico' para outras áreas, como o caso da Fisiologia e Neurofisiologia, ao buscar definir e explicar o mental; e/ou (e) numa definição que implica na falsidade de teses rivais, excluindo-as do debate sobre seu valor de verdade, antes mesmo de ele começar por mera estipulação definicional (como é o caso do pampsiquismo, excluído a priori do debate mente-cérebro pela definição negativa);

(IV) Se o predicado F faz referência, explícita (dimensão física, dimensão temporal, dimensão espacial, partículas físicas, objeto físico, processos biológicos, orgânicos etc.), ou implícita (objetos paradigmaticamente considerados como físicos, objetos macro perceptíveis sensorialmente etc.) a elementos já pressupostos como presentes na propriedade (físico) que necessita ser esclarecida, implica (a) numa definição circular - definir físico como 'dimensão física' ou 'objetos paradigmaticamente considerados físicos' pressupõe aquilo que justamente precisa ser definido, ou (b) numa definição pouco informativa por ser definida de forma sinonímica; ou (c) numa incompatibilidade com as próprias descobertas da Física, tais como eventos existentes não categorizáveis em termos de localização espacial, partículas momentaneamente não identificáveis num momento *t* em nenhum lugar da rede espaço-temporal; ou (d) numa designação de noções novamente da Física, seja ela atual ou futura, tais como "espaço", "tempo", "rede espaço-temporal" e assim por diante, incorrendo nas mesmas limitações das condições (I) e (II).

Até a presente data não existe nenhum consenso sobre quais condições são necessárias e suficientes para adequadamente obtermos uma definição do físico não trivial, não analiticamente verdadeira, não desprovida de conteúdo e passível de permitir o fisicalismo ser uma tese falseável. Qual a relevância de tudo isso? Quais são, então, as implicações lógicas desse problema conceitual para o debate mente-cérebro?

Em primeiro lugar, o fato de que essa questão não consiste meramente de uma disputa terminológica, mas de uma verdadeira questão filosófica. As divergências vão desde quais seriam os critérios apropriados para designação do conceito e tradução do seu conteúdo até se as diferentes categorias de definição

(negativa, positiva, indexadoras etc.) seriam suficientes, não somente para definir o 'físico', ao menos de acordo com uma condição mínima em particular, mas também para superar as severas objeções e obstáculos suscitados pelo problema. O problema do físico, em especial, parece inexoravelmente estar relacionado a outros tópicos de discussão filosófica mais complexos, a saber, a natureza das definições, a natureza das propriedades e relações, assim como a natureza e caracterização dos objetos existentes na realidade. Tradicionais questões e discussões filosóficas estão, implicitamente, associadas ao nosso objeto de investigação, tendo, tal relação, sido negligenciada pelos seus proponentes.

Em segundo lugar, está razoavelmente claro que as diferentes categorias de definição, assim como as diversas condições estipuladas como necessárias pelos fisicalistas para a definição do 'físico' (algumas compatíveis entre si, outras não coextensivas e, por fim, outras incompatíveis) quando contrastadas, indicam no estágio atual de investigação, uma ausência de critério claro quanto à questão, bem como o que venha ser uma formulação adequada do fisicalismo. Se boa parte das críticas forem verdadeiras, a noção de físico seria conceitualmente desprovida de conteúdo, um "vale qualquer coisa" se tiver credenciais científicas. Como consequência, qualquer evento, objeto, estado e/ou propriedade poderia ser categorizada como "físico", desde entidades teorizadas pela Física no presente como desprovidas de massa e de localização espacial determinada até cordas unidimensionais. Logo, não só o fisicalismo seria infalseável, como também trivialmente verdadeiro na ausência de restrições e no estabelecimento de condições mais específicas para a noção de físico. Alguns fisicalistas denominaram de "o problema do corpo" esse difícil obstáculo teórico no contexto da discussão mente-cérebro (MONTERO, 1999).

Em terceiro lugar, o reconhecimento de que o problema da designação do físico implica na própria inexistência de uma tese capaz de ser analisada quanto ao seu valor de verdade. Com efeito, se não temos uma clara definição do 'físico', não temos sequer uma tese fisicalista. E se não temos uma tese fisicalista, a discussão sobre o valor de verdade da tese se torna infrutífera, assim como qualquer tentativa de falseamento da mesma através de evidências empíricas. E levando as derivações ao extremo, como faz Daly, nem mesmo existiria um debate dualismo e fisicalismo, um problema mente-cérebro, caso ambas as teses não consigam lidar com os problemas de definição básicos inerentes às suas caracterizações filosóficas. Corremos o risco, assim, de estarmos discutindo predileções, atitudes, não enunciados e posições filosóficas, conforme já foi sinalizado.

Em quarto lugar, um aprofundamento na análise das diferentes versões fisicalistas, no que se refere às relações por elas estipuladas entre o macrofísico e o microfísico (explicitação da causalidade entre as diferentes ordens dos fenômenos, relações entre propriedades físicas fundamentais e as propriedades mentais, dentre outros), se fazem ainda hoje necessárias. Inclusive, tal problema é investigado, discutido e carente de respostas finais na própria Física contemporânea (BOKULICH, 2008). A análise neste domínio teórico - evidentemente não dissociada do problema conceitual - é crucial no sentido de

evitar a criação indireta de outras aporias, tais como as que têm sido observadas em explicações neurocientíficas atribuindo propriedades psicológicas a áreas cerebrais, estruturas neurofisiológicas de massa encefálica. Tal equívoco em atribuir predicados ao cérebro dotados de sentido somente quando utilizados para referenciar um indivíduo em sua totalidade, uma pessoa, consiste na falácia mereológica (BENNETT & HACKER, 2003). Infelizmente, afirmações do tipo "o cérebro percebe", "um conjunto de neurônios decidiram disparar", dentre tantas outras são mais comuns do que se possa imaginar no campo especializado. Sem perceber, podemos em lugar do fantasma da máquina de Ryle (1949) criar "fantasmas da massa" em tais equívocos lógicos e conceituais. A descrição e exploração específica da relação entre o domínio macro e microfísico são questões estritamente vinculadas ao problema conceitual aqui apresentado, igualmente relevante e de grande impacto nas discussões sobre o problema mente-cérebro.

A presente conjuntura de definições gerais de fisicalismo incompatíveis entre si, propostas divergentes de definições do termo 'físico' utilizadas nas teses, desacordo sobre os critérios de adequação e as condições necessárias para formulação do conceito e a existência de notáveis desafios teóricos como as críticas apresentadas já são suficientes, acredito, para nos fazerem pensar: Por que estamos discutindo no mainstream da Filosofia da Mente o valor de verdade de uma tese que nem mesmo tem conseguido ser formulada de forma adequada? Se não conseguimos definir nem o que é um objeto ou uma propriedade física, não faria sentido perguntar se tal coisa é igual ou diferente de um objeto ou uma propriedade dita mental.

O que estou argumentando é que, na ausência de uma definição do termo 'físico', não é possível existir uma tese geral do fisicalismo, mesmo em meio à também vaga intuição de que a tese deveria afirmar ser tudo o que existe físico, uma vez que dependendo da sua designação, a tese automaticamente pode se transformar em qualquer coisa, inclusive num pampsiquismo, ou numa variação do idealismo. Defendo, portanto, que não temos prestado a devida atenção num passo anterior ao debate mente-cérebro, ou seja, o sério problema da especificação do 'físico', sem a resolução do qual não temos debates concretos à mesa. Em termos comparativos, embora venha sendo discutido nos últimos anos, a questão ainda está num segundo plano em relação a outros tópicos amplamente debatidos em Filosofia da Mente. E na melhor das hipóteses, caso seja possível existir o debate na ausência da própria tese, nossas análises e discussões em nichos específicos circundantes ao problema mente-cérebro têm sido feitas ancoradas em falsas pressuposições de consenso e clareza conceitual. Portanto é imperativo que se busque uma resposta a essa questão antes de dar continuidade a discussões sobre o valor de verdade das divergentes alternativas filosóficas apresentadas como soluções ao problema mente-cérebro.

Igualmente, acredito que a busca por soluções não pode perder de vista alguns questionamentos: Seria essa multiplicidade de teses e definições de físico um sintoma a evidenciar que talvez a ideia, a noção por detrás desse conceito, seja tão difícil de traduzir em palavras quanto os *qualia*? Ou uma mera disputa de preferências teóricas? Seria possível defender a tese fisicalista na ausência de uma definição clara e consensual de 'físico'? Poderia nosso Zeitgeist filosófico estar

equivocado e estar seguindo implicitamente, de forma um tanto quanto acrítica, influências históricas atreladas a um otimismo cientificista? O fato é que não existe uma definição de físico amplamente aceita que não enfrente minimamente uma série de dificuldades teóricas. E isso é um problema tanto para o fisicalista, quanto para o dualista.

Considerações finais

Nossa pesquisa chama atenção para um problema que, embora venha crescentemente sendo reconhecido e analisado por muitos autores, ainda é pouco explorado, de modo geral, nas discussões em torno da natureza da mente humana e do problema mente-cérebro. Consiste numa revisão sistemática sobre o problema da definição do físico na formulação da tese fisicalista, algo até então inexistente no contexto filosófico brasileiro dado à ausência de publicações específicas sobre esse tópico de investigação. Por fim, apresentamos igualmente os obstáculos e aporias envolvidos nesta empreitada teórica, bem como nossa análise lógica-conceitual das implicações para o fisicalismo e o problema mente-cérebro.

É importante reconhecer, igualmente, o fato de que qualquer proposta de definição, de delimitação conceitual da noção de físico pautada na pergunta "O que é o físico no Fisicalismo?", invariavelmente, estará já respondendo implicitamente questões filosóficas basilares e outras, não abordadas anteriormente. Algumas delas são: (a) É possível conhecer aquilo que o termo 'físico' tem como objetivo designar?; (b) É possível definir a noção de 'físico', em termos linguísticos, sem qualquer prejuízo lógico ou distorção do seu *definiens*?; (c) É possível, ao sermos portadores (ainda que em termos cognitivos de primeira pessoa) de uma noção clara da fisicalidade do mundo, sermos capazes de descrevê-lo linguisticamente de forma que possa ser algo intersubjetivamente checável e dotado de um sentido compartilhado?; (d) Pressupondo que existem propriedades no mundo, é possível distinguir propriedades físicas, minimamente, das demais?; e assim por diante.

Tendo isso em vista, sem uma clara definição conceitual, acreditamos ser improvável emergir qualquer resposta definitiva ou solução no nível do valor de verdade das teses. Na ausência de uma formalização clara, da tese fisicalista e de seu conceito central, estamos debatendo em dois níveis: um lógico-empírico, na medida em que se listam argumentos e evidências empíricas a favor e contra; outro conceitual, onde se pressupõem noções intuitivas, porém, mal definidas, de tais propostas. Portanto, argumentamos ser necessário dar um passo atrás na busca pela solução, se existente, para o problema mente-cérebro. Esse passo só pode ser dado tratando e considerando seriamente o problema da definição do físico, bem como todos os problemas ainda mais básicos de Filosofia da Linguagem e da relação entre conceitos e o mundo empírico implicados na questão.

Kim (2005) está parcialmente correto ao afirmar que os esforços atuais em Filosofia da Mente têm sido direcionados predominantemente em como reduzir a mente à dimensão física. Em especial, na busca de encontrar o lugar do "mental" no mundo natural. E o problema é justamente esse. Sustento que esse tipo de

discussão só faz sentido quando temos alguma clareza do que seria essa "base" a qual é reduzida ou vinculada a natureza do mental. Reduzir, ancorar ou vincular, seja por qual tipo de relação for, os fenômenos mentais a um outro domínio, a saber, a dimensão física, cerebral, só faz sentido quando se sabe o que ela é – e do ponto de vista lógico-filosófico, quando se consegue definir sobre o que se está exatamente falando. Talvez esse seja o real "hard problem" da Filosofia da Mente em torno do problema mente-cérebro. Existe um abismo entre o sentido psicológico que automaticamente, cognitivamente, atribuímos ao termo 'físico' e suas diversas acepções e possibilidades de definições, o que favorece a conclusão de que estamos diante de um problema filosófico quando na realidade sequer começamos a compreendê-lo adequadamente. Significar psicologicamente um termo é algo completamente diferente de designar, de forma lógica e mais rigorosa, o conteúdo de uma proposição filosófica.

Ao considerarmos a disputa no nível empírico quanto à natureza do mental, a relevância do nosso objeto de investigação novamente se apresenta como imediata. Dados empíricos não dizem nada por si só, mas precisam ser interpretados. Quanto menos específico e claro é o *framework* teórico-conceitual, maior a probabilidade de qualquer dado ser identificado como evidência favorável por análises *post hoc* e estratégias *ad hoc*. Uma consequência indesejada que antevemos disso tudo é que o debate se torna uma eterna busca por evidências de ambos os lados, sem maior conclusividade, "vencendo" os defensores que tiverem maior poder de influência e persuasão em todas as variáveis de natureza não lógica e não racional (sociais, culturais e econômicas) relacionadas à pesquisa, seja ela filosófica, seja ela científica.

O fato é que não existe uma definição de físico amplamente aceita que não enfrente minimamente uma série considerável de robustas dificuldades teóricas. E de nada adianta uma tese ter influência nos meios acadêmicos se não tem substância filosófica. A temerosa pergunta levantada por Colin McGinn no título de seu artigo "Será que podemos resolver o problema mente e corpo?" remete-nos a Kant em seu primeiro prefácio da Crítica: "A razão humana, num determinado domínio dos seus conhecimentos, possui o singular destino de se ver atormentada por questões, que não pode evitar, pois lhe são impostas pela sua natureza, mas às quais também não pode dar resposta por ultrapassarem completamente as suas possibilidades". Sem embargo, não há razões para acreditarmos que a razão atormentada e insatisfeita deixará de continuar perenemente sua busca por respostas, mesmo sem garantia da possibilidade de factualmente obtê-las.

Referências

ARMSTRONG, David M. *A materialist theory of the mind:* Revised edition. London: Routledge, 1993.

_____. *A World of states of affairs*. Cambridge: Cambridge University Press, 1997.

BEAUREGARD, Mario et al. Manifesto for a post-materialist science. *Explore*, v. 10, n. 5, p. 272-274. 2014.

BENNETT, Max; HACKER, Peter. *Philosophical foundations of neuroscience*. New Jersey: Wiley-Blackwell, 2003.

BISHOP, Robert. The hidden premise in the causal argument for physicalism. *Analysis*, v. 66, p. 44-52, 2006.

BLUMSON, Bem; TANG, Weng. A note on the definition of physicalism. *Thought*, v. 4, p. 10-18, 2015.

BOKULICH, Alisa. *Reexamining the quantum-classical relation:* beyond reductionism and pluralism. Cambridge: Cambridge University Press, 2008.

BOKULICH, Peter. Hempel's dilemma and domains of physics. *Analysis*, v. 7, n. 4, p. 646-651, 2011.

BOURGET, David; CHALMERS, David. What do philosophers believe? *Philosophical Studies*, v. 170, n. 3, p. 465-500, 2014.

BUNGE, Mario. *Matter and mind:* a philosophical inquiry. New York: Springer, 2010.

CARNAP, Rudolf. Psychology in physical language, In: AYER, Alfred (Ed.). *Logical positivism*. New York: The Free Press, 1959. p. 165-198.

CHALMERS, David J. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, v. 2, n. 3, p. 200-219, 1995.

_____. *The conscious mind:* in search of a fundamental theory. Oxford: Oxford University Pres, 1996.

_____. Consciousness and its place in nature. In: STICH, Stephen; WARFIELD, Ted (Eds.). *Blackwell guide to the philosophy of mind.* Oxford: Blackwell, 2002. p.102-142.

CHOMSKY, Noam. *Language and mind*. New York: Harcourt, Brace, Jovanovich, 1972.

_____. Naturalism and dualism in the study of language and mind. *International Journal of Philosophical Studies*, v. 2, n. 2, p. 181-209, 1994.

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

CRANE, Tim; MELLOR, David. There is no question of physicalism. *Mind*, v. 394, p. 185-206, 1990.

CROOK, Seth; GILLETT, Carl. Why physics alone cannot define the 'physical': materialism, metaphysics, and the formulation of physicalism. *Canandian Journal of Philosophy*, v. 31, n. 3, p. 333-359, 2001.

DALY, Chris. What are physical properties? *Pacific Philosophical Quarterly*, v. 79, p. 196-217, 1998.

DENNETT, Daniel. *Kinds of minds:* towards an understanding of consciousness. New York: Basic Books, 1996.

DOVE, Guy. Redefining Physicalism. *Topoi*, v. 37, n. 3, p. 513-522, 2016.

DOWELL, Janice L. The physical: empirical, not metaphysical. *Philosophical Studies*, v. 131, n. 1, p. 25-60, 2006a.

_____. Formulating the thesis of physicalism: An introduction. *Philosophical Studies*, v. 131, n. 1, p. 1-23. 2006b.

FEIGL, Herbert. Physicalism, unity of science and the foundations of psychology. In: COHEN, Robert S. (Ed.). *Inquiries and provocations*: Vienna Circle Collection, v. 14. New York: Springer, 1963. p. 302-341.

_____. *The "mental" and the "physical"*. Minneapolis: University of Minnesota Press, 1967.

FODOR, Jerry A. The mind-body problem. *Scientific American,* v. 244, p. 114-125, 1981.

FRANCESCOTTI, Robert. *Physicalism and the mind*. Dordrecht, Heidelberg, London, New York: Springer, 2014.

GILLET, Carl; LOEWER, Barry. *Physicalism and its discontents*. Cambridge: Cambridge University Press, 2001.

GILLETT, Carl; WITMER, D. Gene. A 'physical' need: physicalism and the via negativa. *Analysis*, v. 61, n. 272, p. 302-309, 2001.

GREYSON, Bruce. The implications of near death experience for a postmaterialist psychology. *American Psychological Association*, v. 2, n. 1, p. 37-45, 2010.

HELLMAN, Geoffrey P.; THOMPSON, Frank W. Physicalism: Ontology, determination, and reduction. *The Journal of Philosophy*, v. 72, n. 17, p. 551-564, 1975.

_____. Physicalist materialism. *Noûs*, v. 11, n. 4, p. 309-345, 1977.

HELLMAN, Geoffrey. Determination and logical truth. *Journal of Philosophy*, v. 82, p. 607-616, 1985.

HEMPEL, Carl G. Reduction: Ontological and linguistic facets. In: MORGENBESSER, Sidney; SUPPES, Patrick; WHITE, Morton (Eds.). *Philosophy, science and method*: essays in honor of Ernest Nagel. New York: St. Martin's Press, 1969. p. 179-199.

_____. Comments on Goodman's "Ways of Worldmaking". *Synthese*, v. 45, n. 2, p. 193-199, 1980.

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

_____. What Mary didn't know. *Journal of Philosophy*, v. 83, n. 5, p. 291-295, 1986.

_____. *From metaphysics to ethics:* A defense of conceptual analysis. Oxford: Oxford University Press, 1998.

JACKSON, Frank. On ensuring that physicalism is not a dual attribute theory in sheep's clothing. *Philosophical Studies*, v. 131, n. 1, p. 227-249, 2006.

JUDISCH, Neal. Why 'non-mental' won't work: On Hempel's dilemma and the characterization of the 'physical'. *Philosophical Studies*, v. 140, n. 3, p. 299-318, 2008.

KIRK, Robert. *Raw feeling:* A philosophical account of the essence of consciousness. Oxford: Oxford University Press, 1996.

_____. *The conceptual link from physical to mental*. Oxford: Oxford University press, 2013.

KIM, Jaegwon. *Mind in a physical world*: An essay on the mind–body problem and mental causation. Cambridge, MA: MIT Press, 1998.

_____. *Physicalism, or something near enough*. Princeton, NJ: Princeton University Press, 2005.

_____. *Philosophy of mind*. 2.ed. Boulder: Westview Press, 2006.

KOONS, Robert C.; BEALER, George. *The waning of materialism*. Oxford: Oxford University Press, 2010.

LEVINE, Joseph. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, v. 64, n. 4, p. 354-361, 1983.

McGINN, Colin. Can we solve the mind-body problem? *Mind*, v. 98, n. 391, p. 349-366, 1989.

MEEHL, Paul; SELLARS, Wilfrid. The concept of emergence. In: FEIGL, Herbert; SCRIVEN, Michael (Eds.). *Minnesota studies in the Philosophy of Science*. Minessota: University of Minessota Press, 1956. p. 239-252.

MELNYK, Andrew. How to keep the 'physical' in physicalism. *The Journal of Philosophy*, v. 94, n. 12, p. 622-637, 1997.

_____. Physicalism unfalsified: Chalmers's inconclusive conceivability argument. In: GILLETT, Carl; LOEWER, Barry M. (Eds.). *Physicalism and its discontents*. Cambridge: Cambridge University Press, 2001. p. 331-349.

_____. *A physicalist manifesto:* thoroughly modern materialism. Cambridge: Cambridge University Press, 2003.

MONTERO, Barbara. The body problem. *Noûs*, v. 33, n. 2, p. 183-200, 1999.

_____. Post-physicalism. *Journal of Consciousness Studies*, v. 8, n. 2, p. 61-80, 2001.

_____. What is the physical? In: BECKERMANN, Ansgar; MCLAUGHLIN, Brian (Eds.). *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press, 2005. p. 173-188.

MONTERO, Barbara. Psysicalism. In: GARVEY, James. *The Continuum Companion to Philosophy of Mind*.1.ed. London: Continuum, 2011. p. 92-101.

MONTERO, Barbara; PAPINEAU, David. A defense of the via negativa argument for physicalism. *Analysis*, v. 65, n. 3, p. 233-237, 2005.

MSIMANG, Phila M. Problems with the 'physical' in physicalism. *South African Journal of Philosophy*, v. 34, n. 3, p. 336-345, 2015.

NAGEL, Thomas. Physicalism. *Philosophical Review*, v. 74, p. 339-356, 1965.

_____. What is it like to be a bat? *Philosophical Review*, v. 83, n. 4, p. 435-450, 1974.

NEURATH, Otto. Physicalism. In: COHEN, Robert: NEURATH, Otto (Eds.). *Philosophical Papers 1913–1946*. Dordrecht: D. Reidel, 1931a. p. 52-57.

_____. Physicalism: the philosophy of the Vienna Circle. In: COHEN, Robert, S.; NEURATH, Otto (Eds.). *Philosophical Papers 1913–1946*. Dordrecht: D. Reidel, 1931b. p. 48-51.

NEY, Alyssa. Defining physicalism. *Philosophy Compass*, v. 3, n. 5, p. 1033-1048, 2008.

NIMTZ, Christian; SCHUTTE, Michael. On physicalism, physical properties, and panpsychism. *Dialectica*, v. 57, n. 4, p. 413-422, 2003.

PAPINEAU, David. The reason why: response to Crane. *Analysis*, v. 51, n. 1, 37-40, 1991.

_____. *Philosophical naturalism.* Oxford: Blackwell, 1993.

_____. The rise of physicalism. In: GILLETT, Carl; LOEWER, Barry (Eds.). *Physicalism and its discontents*. Cambridge: Cambridge University Press, 2001.

PARNIA, Sam. Do reports of consciousness during cardiac arrest hold the key to discovering the nature of consciousness? *Medical Hypotheses*, v. 69, n. 4, p. 933-937, 2007.

PETTIT, Philip. A definition of physicalism. *Analysis*, v. 53, n. 4, p. 213-223, 1993.

PLACE, Ullin. Is consciousness a brain process? *British Journal of Psychology*, v. 47, n. 1, p. 44-50, 1956.

POLAND, Jeffrey. *Physicalism:* the philosophical foundations. Oxford: Clarendon Press, 1994.

_____. Chomsky's challenge to physicalism. In: ANTONY, Louise; HORNSTEIN, Norbert (Eds.). *Chomsky and his critics*. Hoboken: Blackwell, 2003. p. 29-48.

POPPER, Karl; ECCLES, John. *The self and its brain*. Berlin, Heidelberg: Springer, 1977.

PUTNAM, Hilary. Psychological predicates. In: CAPITAN, William; MERRILL, Daniel (Eds.). *Art, mind, and religion*. Pittsburgh: University of Pittsburgh Press, 1967. p. 37-48.

QUINE, Willard V. The scope and language of science. In: QUINE, Willard V. *The ways of paradox and other essays.* Cambridge, MA: Harvard University Press, 1954. p. 228-245.

ROBINSON, William. The hardness of the hard problem. *Journal of Consciousness Studies*, v. 3, n. 1, p. 14-25, 1996.

RYLE, Gilbert. *The concept of Mind.* New York: University Paperbacks; Barnes & Noble, 1949.

SCHAFFER, Jonathan. Is there a fundamental level? *Noûs*, v. 37, n. 3, p. 498-517, 2003.

SCHEFFLER, Israel. The new dualism: psychological and physical terms. *The Journal of Philosophy,* v. 47, p. 737-752, 1950.

SCHIFFER, Stephen. Physicalism. *Philosophical Perspectives*, v. 4, p. 153-185, 1990.

SEARLE, John. *The rediscovery of the mind*. Cambridge, MA: MIT Press. 1992.

SHELDRAKE, Rupert. Setting science free from materialism. *Explore*, v. 9, n. 4, p. 211-218, 2013.

SMART, John J. Sensations and brain processes. *The Philosophical Review*, v. 68, n. 2, p. 141-156, 1959.

_____. Materialism. *The Journal of Philosophy*, v. 60, n. 22, p. 651-662, 1963.

_____. The content of physicalism. *Philosophical Quarterly*, v. 28, n. 113, p. 339-341, 1978.

_____. Physicalism and emergence. *Neuroscience,* v. 6, n. 2, p. 109-113, 1981.

SPURRETT, David. What physical properties are. *Pacific Philosophical Quarterly*, v. 82, n. 2, p. 201-225, 2001.

_____. Physicalism as an empirical hypothesis. *Synthese*, v. 194, n. 9, p. 3347-3360, 2015.

SPURRETT, David; PAPINEAU, David. A note on the completeness of "physics". *Analysis*, v. 59, n. 1, p. 25-29, 1999.

STOLJAR, Daniel. Two conceptions of the physical. *Philosophy and Phenomenological Research*, v. 62, n. 2, p. 253-281, 2001a.

_____. The conceivability argument and two conceptions of the physical. *Philosophical Perspectives*, v. 15, p. 393-413, 2001b.

_____. *Physicalism*. London: Routledge, 2010.

_____. Physicalism. In: ZALTA, Edward N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Disponível em: <https://plato.stanford.edu/entries/physicalism/>. Acesso em: 07 jun. 2016.

STRAWSON, Galen. Realistic monism: why physicalism entails panpsychism. *Journal of Consciousness Studies*, v. 13, n. 10/11, p. 53-74, 2006.

STURGEON, Scott. Physicalism and overdetermination. *Mind*, v. 107, n. 426, p. 411-432, 1998.

SWINBURNE, Richard. *The evolution of the soul*. Oxford: Oxford University Press, 1986.

VAN FRAASSEN, Bas. Science, materialism, and false consciousness. In: KVANVIG, Jonathan (Ed.). *Warrant in contemporary epistemology:* essays in honor of Alvin Plantinga's theory of knowledge. Lanham: Rowman Littlefield, 1996. p. 149-182.

VAN LOMMEL, Pim. Non-local consciousness: A concept based on scientific research on near-death experiences during cardiac arrest. *Journal of Consciousness Studies*, v. 20, n. 1/2, p. 07-48, 2013.

VICENTE, Agustín. Current physics and 'the Physical'. *British Journal for the Philosophy of Science*, v. 62, n. 2, p. 393-416, 2011.

WILSON, Jessica. On characterizing the physical. *Philosophical Studies*, v. 131, n. 1, p. 61-99, 2006.

WITMER, D. Gene. Physicality for physicalists. *Topoi*, v. 37, n. 3, p. 457-472, 2016.

ZILIO, Diego. Fisicalismo na filosofia da mente: definição, estratégias e problemas. *Ciências & Cognição,* v. 15, n. 1, p. 217-240, 2010.

WORLEY, Sara. Physicalism and the via negativa. *Philosophical Studies*, v. 131, n. 1, p. 101-126, 2006.

Recebido em: 06-03-2019

Aceito para publicação em: 19-08-19

A DISCUSSÃO EM TORNO DA "PARTE II" (TS 234) DAS *INVESTIGAÇÕES FILOSÓFICAS* DE WITTGENSTEIN

THE QUARREL ABOUT "PART II" (TS 234) OF WITTGENSTEIN'S PHILOSOPHICAL INVESTIGATIONS

FILICIO MULINARI[1]
Universidade Federal de São Paulo (UNIFESP) – Brasil
filicio@gmail.com

RESUMO: O artigo tem dois propósitos centrais: primeiramente, expor a discussão sobre a inclusão do TS 234 (anteriormente conhecida como "Parte II") nas *Investigações Filosóficas* de Wittgenstein. Para atingir esse primeiro objetivo, iremos analisar as justificativas de inclusão dadas por Anscombe (1953) e Rhees (1953, 1996) na primeira edição do livro, as críticas feitas por von Wright (1982, 1992) e, por fim, os motivos que levaram Hacker e Schulte, editores da quarta edição do livro, a considerar o TS 234 como parte de um trabalho distinto das *Investigações Filosóficas*. O segundo objetivo do artigo é mostrar uma leitura particular sobre os escritos pós-1945 (incluindo o TS 234), a qual afirma que tais anotações não representam 'partidas para novas direções', como von Wright (1982), Hacker (2013) e Moyal-Sharrock (2004) alegam, mas são uma continuação das questões iniciadas ainda nas *Investigações Filosóficas*. Para este segundo objetivo, tomaremos como base a leitura exposta por Nuno Venturinha (2007) em seu artigo "Against a third Wittgenstein" e iremos nos fundamentar em menções do próprio Wittgenstein que indicam que as *Investigações Filosóficas* ainda estavam no horizonte de pensamento do filósofo em seus escritos pós-1945. Ao fim, concluiremos que tanto a leitura proposta por Venturinha – de "uma continuidade" dos escritos – quanto as leituras que apregoam que os escritos sobre filosofia da psicologia são "partidas para novas direções" possuem limitações e problemas graves, o que mostra que a questão em torno do *locus* da Parte II das *Investigações* e dos problemas ali inseridos continua sendo ainda hoje um capítulo polêmico das leituras Wittgensteinianas.

PALAVRAS-CHAVE: Wittgenstein. Filosofia da psicologia. TS 234. Conceitos psicológicos.

ABSTRACT: This paper has two central purposes: first, to discuss the inclusion of TS 234 (formerly known as "Part II") in Wittgenstein's *Philosophical Investigations*. In order to reach this first objective, we will analyse the justifications for inclusion given by Anscombe (1953) and Rhees (1953, 1996) in the first edition of the book, the criticisms made by von Wright (1982, 1992) and, finally, the reasons which led Hacker and Schulte, editors of the fourth edition of the book, to consider TS 234 part of a work distinct from the *Philosophical Investigations*. The second objective of the paper is to argue for a certain reading of the post-1945 writings (including TS 234), which claims that such annotations do not represent 'departures in new directions', such as von Wright (1982), Hacker (2013), and Moyal-Sharrock (2004) argue, but they are a continuation of the questions initiated in the *Philosophical Investigations*. For this second objective, we will discuss the reading proposed by Nuno Venturinha (2007) in his article "Against a third Wittgenstein", taking as our starting-point some writings of Wittgenstein that indicate that the *Philosophical Investigations* was still on the horizon in his post-1945 writings. At the end, we will conclude that both the reading proposed by Venturinha - of "a continuity" in the writings - and the readings that support that the writings on philosophy of psychology are "departures in new directions" have limitations and serious problems, which shows that the question about the *locus* of Part II of the *Investigations* and the problems therein still remain a controversial chapter in Wittgenstein studies.

KEYWORDS: *Wittgenstein. Philosophy of psychology. TS 234. Psychological concepts.*

[1] Doutor em Filosofia pela Universidade Federal de São Paulo (Unifesp).

Introdução

É de conhecimento público que o texto das *Investigações Filosóficas* (*PU*) não estava terminado à época da morte de Wittgenstein, em 1951. Por este motivo, coube então aos editores a edição para a publicação póstuma. A primeira edição das *PU,* publicada em 1953, trazia o livro dividido em duas partes intituladas como "Parte I" (*Teil I*), que possuía 693 parágrafos numerados, e "Parte II" (*Teil II*), dividida em quatorze seções referenciadas por algarismos romanos.

A divisão do texto das *PU* em "Parte I" e "Parte II" continuou presente nas duas edições posteriores da obra e serviu de modelo para várias traduções durante muitos anos.[2] Para se ter uma ideia do tamanho da influência, basta notar que todas as edições brasileiras das *Investigações Filosóficas* trazem o texto dividido em duas partes.[3] Entretanto, embora a divisão do texto em partes distintas já fosse motivo de debates há décadas, a discussão tomou novos rumos com a publicação da 4ª edição inglesa das *PU*, publicada em 2009. Peter Hacker e Joachim Schulte, editores da 4ª edição, decidiram eliminar a divisão da obra e, assim, apenas a antiga Parte I recebeu o título de *Investigações Filosóficas*. A 'Parte II' foi renomeada para *Filosofia da Psicologia – Um Fragmento,* entendida agora como sendo um fragmento de um trabalho distinto, um 'trabalho em andamento' [*work in progress*], nos termos dos editores).

Dada a influência das antigas edições, o que propomos na sequência é apresentar uma exposição do debate acerca da inclusão/exclusão da antiga Parte II (TS 234) nas *PU*, desde a primeira edição (1953) até a edição feita por Hacker e Schulte (2009), versão mais recente e revisada do texto. Após isso, iremos expor a leitura de Venturinha (2007) que afirma que os escritos de Wittgenstein pós-1945 (incluindo a antiga Parte II) não tratam de temas distintos, mas são uma sequência do debate iniciado ainda nas *Investigações Filosóficas.* O objetivo, ao final, é mostrar que tanto a decisão de Hacker e Schulte de retirar o TS-234 das *Investigações* quanto a leitura continuísta de Venturinha a respeito do texto possuem incoerências teóricas e limitações documentais, o que nos mostra que a querela em torno do lugar do TS 234 no pensamento de Wittgenstein continua sendo um tópico atual e importante nos estudos das anotações do filósofo.

[2] As duas edições de referência que mencionamos são, respectivamente, a 2ª Ed. publicada pela Blackwell Publishers (1958), e a 3ª edição, publicada em 2003 pela Blackwell Publishers. De acordo com Hacker (2009, prefácio, p. viii), a segunda edição trazia pequenas correções de pontuação e grafia no texto alemão, além de um grande número de pequenas mudanças e 28 mudanças significativas no texto em inglês. A terceira edição, publicada na ocasião do 50º aniversário da primeira publicação, trazia um pequeno número de alterações à tradução de Anscombe (tradutora da primeira versão).

[3] Tomamos como referência brasileira as nove edições das IF publicadas pela Editora Vozes, sendo a última datada de 2014, e as versões pertencentes à *Coleção Os Pensadores*, publicadas pela Editora Abril e pela Editora Nova Cultural entre os anos de 1973 e 2004, que findou sua série de publicações das *IF* na 7ª Edição, publicada no ano de 1999.

1 A HISTÓRIA CONTURBADA DA "PARTE II" DAS PU: VON WRIGHT E SUA INFLUÊNCIA

Após a morte de Ludwig Wittgenstein, em 29 de abril de 1951, uma grande quantidade de escritos não publicados ficou sob os cuidados de três de seus antigos alunos: Rush Rhees, Elizabeth Anscombe e Georg Henrik von Wright. Esse espólio filosófico póstumo de Wittgenstein, conhecido pelo termo alemão *Nachlass,* foi prontamente trabalhado por esses alunos que começaram a mapear o que poderia ser material para edição e publicação.

Inicialmente, Rhees e Anscombe - editores da primeira edição das *PU,* publicada em 1953 - decidiram publicar em formato de livro os textos datilografados que consideravam estarem mais próximos de um livro acabado.[4] É nesse sentido vem à tona a publicação das *Investigações Filosóficas,* utilizando como base o TS 227, adicionando também o TS 234 (que corresponde a antiga "Parte II" das *PU*).

A escolha dos editores para a publicação do TS 227 junto com o TS 234 não foi isenta de questionamentos por parte da comunidade acadêmica ao longo dos anos. Era de conhecimento dos editores que Wittgenstein havia trabalhado nas *PU* durante anos, inclusive tendo o próprio filósofo escrito ao menos cinco versões prévias (*drafts*) do texto.[5] É com certo grau de certeza, por exemplo, que podemos dizer que os primeiros 189 parágrafos das *PU* já estavam prontos ao fim da década de 1930. Mais que isso, também podemos dizer que há 'blocos de texto' pertencentes as *PU* e que foram escritos em momentos diferentes. Assim, sabemos que os parágrafos 1-410 já estavam prontos entre os anos de 1943 e 1945 (versão intermediária) e o restante das anotações da Parte I (§§411-693) só vieram a ser devidamente finalizadas no ano letivo de 1945/46. No que diz respeito à antiga Parte II, ela foi redigida entre maio de 1946 e maio de 1949 e corresponde ao manuscrito MS 144. Esse manuscrito, por sua vez, foi datilografado e deu origem ao TS 234.[6]

Os primeiros a refletirem sobre uma suposta incongruência em publicar a Parte II (TS 234) no texto final das *PU* foram Anscombe e Rhees, editores da primeira edição. Na 'Nota dos Editores', eles ponderam que se Wittgenstein tivesse vivido, ele provavelmente teria alterado boa parte dos últimos parágrafos da Parte I e incluído alguns conteúdos da Parte II em seu lugar:

---

[4] Os escritos de Wittgenstein são tradicionalmente catalogados enquanto manuscritos (MS) e textos datilografados (em inglês, Typescript ou TS). Para referência aos anos em que foram escritos os materiais publicados postumamente, tomou-se como referência o trabalho catalográfico desenvolvido por Biggs e Pichler (1993).

[5] As cinco versões prévias mencionadas aqui estão catalogadas na Edição Crítico-Genética das Investigações Filosóficas realizada por Joachim Schute, Heikki Nyman, Eike von Savigny e George H. von Wright (Editora Suhrkamp, 2001). Elas recebem o nome de *Urfassung* (MS 142), *Frühfassung* (TS 225, TS 220, TS 221), *Bearbeitete Frühfassung* (TS 239), *Zwischenfassung* e por fim, ao *Spätfassung* (TS 227).

[6] "O conjunto de anotações presentes na Parte II foram selecionadas por Wittgenstein dos manuscritos escritos entre maio de 1946 e maio de 1949. Mais da metade deles datam de outubro de 1948 a março de 1949" (von WRIGHT, 1992, p. 182).

> O que se apresenta neste volume como Parte I estava pronto desde 1945. A Parte II surgiu entre 1947 e 1949. Se o próprio Wittgenstein tivesse publicado sua obra, teria deixado de lado grande parte daquilo que agora constitui aproximadamente as últimas trinta páginas da Parte I e, no lugar delas, teria inserido o conteúdo da Parte II, com acréscimo e novo material. […] Somos responsáveis pela ordenação dos últimos fragmentos da Parte II no seu lugar atual. (ANSCOMBE; RHEES, 1953, *nota dos editores*).[7]

Mesmo Anscombe e Rhees tendo o cuidado de explicitar as dificuldades relacionadas à Parte II, será George Henrik von Wright que aprofundará a questão e lançará questionamentos mais pontuais a respeito da decisão editorial em dividir as *IF* em duas partes. Em seu artigo "The Troubled History of Part II of the Investigations" (1992), von Wright explicita alguns questionamentos entre o texto datilografado e aquilo que era considerado como parte integrante das *PU*, salientando que o próprio título "Parte II" fora uma invenção dos editores (von WRIGHT, 1992, p. 181).

> Não está claro para mim por que razão os editores utilizaram dois datiloscritos para imprimir o que eles consideraram como uma só e a mesma obra. Só o primeiro desses dois textos escritos à máquina (TS 227 no catálogo dos escritos póstumos de Wittgenstein) trazia no cabeçalho, vindo do próprio autor, "Investigações Filosóficas". O segundo datiloscrito não tinha nenhum título. As designações "Parte I" e "Parte II" são provenientes dos editores. (von WRIGHT, 2001, p. 7).

O artigo de von Wright teve uma influência direta na editoração da 4ª Edição das *PU,* sobretudo no que diz respeito a decisão de excluir a divisão do livro em duas partes. Vale lembrar que essa decisão foi, segundo os editores, a mudança editorial mais importante feita na obra (HACKER; SCHULTE, 2009, p. ix). No texto editorial explicativo, os editores contestam a nota de Anscombe e Rhees que afirma que "[...] se o próprio Wittgenstein tivesse publicado sua obra, ele teria suprimido boa parte das últimas trinta páginas da Parte I e, no lugar delas, inserido o conteúdo da Parte II".[8] Para Hacker e Schulte, não haveria nenhuma evidência de que o filósofo realmente havia pensado em uma publicação conjunta da antiga Parte I junto com a Parte II:

> Não há nenhuma evidência escrita no *Nachlass* de Wittgenstein ou correspondência que sugira que o MS 144 tenha sido feito para coletar materiais que seriam incorporados nas *Investigações Filosóficas*. Nem há qualquer indício que ele tenha pensando em suprimir "uma boa parte daquilo que constitui aproximadamente as últimas trinta páginas da Parte I". (HACKER; SCHULTE, 2009, p. xxii).

[7] Todas as traduções no decorrer do texto são do autor.
[8] O texto editorial explicativo recebeu o nome de *The Text of the Philosophische Untershuchungen* e vem logo após o prefácio editorial, na própria 4.a edição das *PU.*

A carência de evidências escritas, como notas ou correspondências que indicassem o desejo de Wittgenstein em publicar a antiga "Parte II" junto com a "Parte I" das *PU,* constitui um forte elemento para que Hacker e Schulte não vejam com bons olhos a decisão de publicar o TS 234 como sendo uma 'segunda parte' das *PU*. Noutras palavras, para eles a inclusão dessa parte seria uma decisão meramente editorial (e errônea), não de Wittgenstein.

> Entre os artigos de Wittgenstein, os editores acharam um texto datilografado [TS 234] fundamentado no manuscrito MS 144. Era uma coletânea de 372 anotações não numeradas selecionadas em sua maioria dos manuscritos redigidos entre maio de 1946 e maio de 1949. Anscombe e Rhees **decidiram** que este texto datilografado era parte do mesmo livro dos 693 parágrafos numerados que eles chamaram de "Parte I" (HACKER; SCHULTE, 2009, p. xxi, grifo-nosso).

Para fundamentar ainda mais sua posição contrária à inclusão do TS 234 nas *PU*, Hacker e Schulte, além de se apoiarem na carência de evidência escrita, fazem uso do trabalho de von Wright, sobretudo do questionamento que ele faz a respeito da data na qual Wittgenstein teria se encontrado com os primeiros editores.

Segundo Hacker e Schulte - fundamentados no artigo de von Wright -, Wittgenstein teria se encontrado com Anscombe e Rhees em dezembro de 1948, em Dublin. Ali o filósofo teria discutido suas intenções sobre as *PU* aos futuros editores, uma vez que supostamente eles não teriam se encontrado depois dessa data.[9] O problema residiria no fato de que, naquele momento, Wittgenstein ainda não teria compilado o MS 144 e nem ditado o TS 234. Por isso, ainda que o filósofo tivesse a intenção de 'revisar as últimas 30 páginas do livro', como afirmaram Anscombe e Rhees, o fato é que ele nunca chegou a efetuar essas revisões e, além disso, nem sequer havia ainda o texto datilografado da Parte II (TS 234).

> Naquele momento [do encontro de Wittgenstein com Anscombe e Rhees em Dublin], a maior parte daquilo que foi coletado no MS 144 tinha sido escrito em manuscritos muito mais extensos (MS 137 e MS 138). Mas nem o MS 144 nem, é claro, o subsequente texto datilografado TS 234 haviam sido compilados. Pode muito bem ter sido que nesse momento Wittgenstein contemplava revisar as últimas 30 páginas de seu livro e pensado em usar alguns dos vastos materiais que ele havia escrito desde 1946 nesse processo. Porém, ele nunca realizou nenhuma dessas intenções (HACKER; SCHULTE, 2009, p. xxii).

O próprio Hacker, em um trabalho individual posterior à 4ª edição das *PU*, sustenta a mesma visão sobre a impossibilidade de se saber o que Wittgenstein tinha em mente com a Parte II:

[9] Devemos analisar especialmente a observação de von Wright que diz respeito a Parte II: "Por tudo que pude verificar, Wittgenstein não falou sobre seus planos aos futuros editores das *Investigações* após ele ter deixado Dublin em 1949" (von WRIGHT, 1992, p. 187).

> [...] É notável que a conversa na qual Wittgenstein contou a Rhees e Anscombe que ele tinha a intenção de suprimir parte das últimas 30 páginas das *Investigações* e trabalhar 'e inserir naquela parte aquilo que é a Parte II, junto com outros materiais' (PU, nota dos editores)' aconteceu em Dublin, em dezembro de 1948, *antes* dele ter escrito o MS 144, e seis meses antes dele tê-lo ditado. Assim sendo, é impossível ter certeza sobre o que Wittgenstein tinha em mente naquele momento (HACKER, 2013, p. 80 - nota 3).

Segundo Hacker e Schulte, a única coisa dessa querela que é conhecida com certeza é que o MS 144 foi datilografado no final de junho e início de julho, em 1949, dando origem ao TS 234, que seria usado depois para publicação da "Parte II". Ou seja, o manuscrito só foi ditado meses após o encontro com Anscombe e Rhees. Por sinal, é bem provável que Wittgenstein tenha datilografado esse manuscrito com a intenção mostrá-lo a seu amigo Norman Malcolm, na visita que viria a fazer a ele no final de julho de 1949. O próprio Malcolm faz uma consideração bastante reveladora sobre as intenções de Wittgenstein com as *PU* à época:

> Em uma de nossas caminhadas, Wittgenstein disse que se ele tivesse dinheiro, ele iria mimeografar e distribuir seu livro (a Parte I das *Investigações*) a seus amigos. Ele disse que o livro não estava completamente acabado, mas que ele não pensava que poderia terminá-lo em vida. O plano dele era colocar em parênteses, após as anotações, expressões de desaprovação, como "isto não está muito certo" ou "isto é suspeito". Ele gostaria de colocar seu livro nas mãos de amigos, mas levá-lo a uma editora estava fora de cogitação (MALCOLM, 1984, p. 75).

Para Hacker e Schulte, o comentário de Malcolm fortalece ainda mais a ideia de que Wittgenstein não tinha em mente a ideia de reescrever radicalmente as últimas 30 páginas do livro, tal como afirmam Anscombe e Rhees. Afinal, segundo Malcolm, Wittgenstein já teria desistido naquele momento de ter uma versão mais acabada de seu livro em vida e faria no máximo anotações entre parênteses. Logo, independente das intenções de Wittgenstein, Hacker e Schulte concluem que o fato é que o mais próximo que o filósofo chegou de terminar as *PU* foi com o conjunto de parágrafos 1-693.

A incongruência da suposta data do último encontro entre Anscombe e Rhees com Wittgenstein, somada à declaração de Malcolm e a falta de evidências escritas, indica que a antiga "Parte II" nada mais seria que um rearranjo de um conjunto de parágrafos escritos entre 1946 e 1949, não tendo vínculo direto com o trabalho das *Investigações Filosóficas*. Isso justificaria a decisão dos editores da 4ª Ed. de retirar tal divisão, dando a entender que as *PU* seriam referentes somente à antiga Parte I. Entretanto, a decisão de Hacker e Schulte também é passível de críticas – e é a elas que nos debruçaremos na sequência.

2 Uma crítica à leitura de Hacker e Schulte

Em seu artigo que fundamenta a visão de Hacker e Schulte, von Wright relembra que em junho de 1949, em uma carta a Norman Malcolm, Wittgenstein dizia que estava pretendendo ditar alguns trabalhos nas próximas duas semanas enquanto estivesse em Cambridge. Após isso, Wittgenstein levou consigo o texto datilografado mencionado na carta aos EUA e entregou a Malcolm em julho de 1949. Além disso, von Wright relembra que Anscombe confirmou a ele próprio que Wittgenstein estava ditando alguns trabalhos naquele momento – muito provavelmente o manuscrito MS 144. E mais: o próprio von Wright lembra de ter discutido o MS 144 (ou o TS 234) com Wittgenstein nesse período que antecedeu a viagem aos EUA, quando o filósofo ainda estava em Cambridge. Entretanto, von Wright questiona se Anscombe e Rhees tinham conhecimento das intenções de publicação de Wittgenstein à época. Ele recorda – tal como assumido por Hacker e Schulte – que Anscombe e Rhees haviam encontrado Wittgenstein em Dublin, no ano de 1948. Contudo, naquele momento Wittgenstein não havia redigido o MS 144 nem datilografado o TS 234, tal como afirmado anteriormente.

> Quando Wittgenstein disse a Anscombe e Rhees sobre seus planos? Ambos visitaram Wittgenstein em Dublin em Dezembro de 1948. Neste momento a maior parte da Parte II tinha sido escrita – mas nem o manuscrito MS 144 e nem, consequentemente, qualquer texto datilografado dessa parte do livro existia. [...] Por tudo que pude verificar, Wittgenstein não falou sobre seus planos aos futuros editores das *Investigações* após ele ter deixado Dublin em 1949 (von WRIGHT, 1992, p. 186-187).

De acordo com a citação acima, parece pouco provável que Wittgenstein tenha dito aos primeiros editores das *PU* que ele gostaria de ter 'alterado as últimas 30 páginas', menos provável ainda é de que naquela ocasião ele tenha afirmado qualquer coisa sobre a "Parte II". Afinal, ela nem sequer havia sido datilografada. Porém, alguns questionamentos contrários às afirmações de von Wright podem ser feitos.

Recentemente foram publicadas algumas correspondências dos primeiros editores que dão ainda mais corpo à discussão sobre a inserção do TS 234 nas *PU*. Em uma carta a von Wright de 1972, Rush Rhees fala sobre as intenções de Wittgenstein a respeito da "Parte II".[10] Nela, Rhees afirma categoricamente que Wittgenstein tinha a intenção de incluir partes dos seus recentes manuscritos nas *PU*, ainda que não indicasse quais partes do manuscrito de fato deveriam ser incluídas.

> A principal 'revisão' na qual ele [Wittgenstein] estava trabalhando no fim de 1948 e no início de 1949 foi na Parte II (como a chamamos). Ele estava trabalhando muito nisso quando eu o visitei em Dublin nas férias de Natal de 1948/1949 (aproximadamente entre 20 de dezembro e 10 de janeiro). Ele falou sobre as partes que ele havia concluído, [sic] leu algumas delas para mim. Mas ele

[10] A carta de Rhees a von Wright é datada de 10 de Agosto de 1972. Ela se encontra no von Wright e Wittgenstein Archives da Universidade de Helsinnki (WWA).

> não explicou quais partes do manuscrito "Parte I" deveriam ser substituídas (RHEES apud ERBACHER, 2015, p. 171).

Além da menção direta de Rhees a respeito das intenções de Wittgenstein, outro fator que ajuda a colocar em xeque a leitura de von Wright é o fato de que Elizabeth Anscombe teria se encontrado com Wittgenstein em Viena um ano *após* a visita em Dublin. O testemunho é de Peter Geach, ex-aluno de Wittgenstein e marido de Anscombe.

> No final de 1949 Wittgenstein foi a Viena e permaneceu por alguns meses; sua estadia coincidiu com uma longa visita de Elizabeth [Anscombe] a seus amigos vienenses; ela já havia se comprometido a traduzir as *Investigações* e buscado se preparar para a tarefa com um bom conhecimento do alemão vienense (GEACH, 1988, xiii).

É interessante notar ainda que o encontro entre Wittgenstein e Anscombe mencionado na citação acima aconteceu após a visita do filósofo a Malcolm, tendo o datiloscrito da Parte II (TS 234) já sido finalizado. Mais que isso, é evidente que Anscombe foi à Viena e se encontrou com Wittgenstein com a finalidade de discutir sobre a edição das *PU*. O próprio Wittgenstein confirma o encontro com Anscombe em uma carta a Malcolm, datada de 12 de fevereiro de 1950.[11] Logo, é bem provável que Anscombe tivesse ao menos em parte conhecimento das pretensões futuras de Wittgenstein sobre o texto, inclusive sobre a relação do TS 234 com as *PU*.

O conteúdo das cartas e os testemunhos aqui mencionados vão em direção contrária às conclusões feitas por von Wright em seu artigo, ou, ao menos, à interpretação dada por Hacker e Schulte sobre o TS 234. Se formos mais atentos, torna-se difícil até mesmo vincular a decisão de Hacker e Schulte em excluir a Parte II na 4ª Edição com as próprias conclusões de von Wright. Afinal, ao tomarmos as anotações de von Wright de forma conjunta, uma interpretação mais plausível seria aquela que indica que ele simplesmente não poderia afirmar *com certeza* sobre a forma final do novo material que Wittgenstein gostaria de inserir na Parte I, mas não que *não haja* nada a ser inserido ali. Embora a Parte II não fosse um 'produto acabado', von Wright não parece ter dúvida de que não haveria melhor lugar para inseri-la do que junto à primeira parte das *IF*:

> Parece a mim, em um fundamento intrínseco, certo que Wittgenstein considerava o texto datilografado da Parte II como um produto mais acabado que outros textos que ele havia ditado após o texto da Parte I. Não posso ver nenhuma boa razão para que os editores, Anscombe e Rhees, não tivessem o publicado junto com a Parte I (von WRIGHT, 1992, p. 188).

Paralelo ao equívoco sobre aos encontros dos editores com Wittgenstein – que, como se nota, não aconteceu por último em Dublin, ao menos no que diz respeito à Anscombe -, também devemos mencionar que von Wright admite que 'nunca havia discutido sobre as intenções de publicação de Wittgenstein com ele

[11] Querido Norman, [...] Eu pretendo ficar aqui mais um mês. Estou muito bem de saúde. Tanto é que eu tive uma discussão muito boa com Sr.ª Anscombe há alguns dias (MALCOLM, 2001, p. 126).

próprio' e, ainda, que 'soube que era um de seus executores literários [*Literary Executors*] somente *após* a morte do filósofo' (vide von WRIGHT, 1992, p. 188). Em contrapartida, Rhees pontua:

> Um pouco antes de sua morte, Wittgenstein estava falando comigo sobre o trabalho de edição de seus manuscritos. Isto ficou constantemente em minha cabeça, e isto era muito especial. Ele disse: "Eu confio absolutamente em você e na Sr.ª Anscombe" (RHEES, 1996, p. 56).

Ainda que não seja o caso de entrarmos em uma discussão interminável sobre qual dos executores literários estaria mais certo no que diz respeito à edição das *PU*, é um fato dado que Wittgenstein não tratou da edição de seus escritos com ninguém além de Anscombe e Rhees. Cabe reafirmar aqui a conclusão de von Wright, na qual ele afirma: "não posso ver nenhuma boa razão para que os editores não tivessem publicado [a Parte II] junto com a Parte I" (1992, p. 188). Tendemos a concordar com von Wright, ainda que isso em certa medida nos distancie da decisão editorial de Peter Hacker e Joachim Schulte. Mais que isso, também fazemos coro com outra conclusão de von Wright, na qual ele afirma: "As *Investigações Filosóficas* permanecem um *torso*. Isso [a edição do livro] pode não ser satisfatório. Mas sobre isso ninguém pode culpar os editores, que fizeram seu melhor para apresentar o livro ao mundo (von WRIGHT, 1992, p, 188)."

Compartilhamos ainda mais dessa conclusão específica de von Wright mencionada acima por dois motivos: primeiramente, por reconhecer o mérito e esforço de Anscombe e Rhees na edição das *PU*; segundo, por abrir margem para a ideia de que a antiga "Parte II" não deve ser entendida como um 'trabalho paralelo' alheio às *PU*, mas sim como seu desdobramento. É exatamente para esse segundo tópico que nos voltamos agora.

3 Os escritos sobre filosofia da psicologia e a Parte II: uma tese da continuidade das *PU*

Alguns elementos biográficos são bem consistentes com a ideia de que Wittgenstein continuou a trabalhar no texto das *Investigações Filosóficas* até os últimos anos de sua vida. Entretanto, essa tese vai de encontro com a visão bastante difundida de que a antiga Parte I das *PU* seria um 'projeto acabado' e, consequentemente, a "Parte II" não deveria ser incluída no mesmo livro. Segundo esse raciocínio, a Parte II e todo o seu conteúdo pertenceria a outro eixo temático *distinto* daquele das *PU*. Novamente, von Wright (1982) é um dos pioneiros dessa visão:

> [eu me] inclino a aceitar a opinião de que a Parte I das *Investigações Filosóficas* é um trabalho completo e que os escritos de Wittgenstein de 1946 em diante representam em certa medida partidas para *novas* direções (von WRIGHT, 1982, p. 136).

A afirmação de von Wright tomou força ao longo dos anos. Como exemplo, temos a leitura de Hacker, feita quase 20 anos depois, que revela não só proximidade, mas também uma efetiva influência da conclusão de von Wright em sua leitura:

> Pode muito bem ser que as tentativas de realocar as anotações da Parte II na Parte I tenham falhado. Certamente é difícil ver como tanto material, em particular a longa discussão sobre 'percepção de aspecto', poderia ter sido incorporada no texto existente sem grandes mudanças à estrutura do argumento. Então pode ser que, caso Wittgenstein tivesse vivido e continuado suas últimas investigações sobre filosofia da psicologia (RPPI e II, LWI e II), ele teria incorporado esta grande quantidade de material adicional em um volume separado sobre filosofia da psicologia. Professor G.H. Wright escreveu que ele se inclina *a aceitar a opinião de que a Parte I das Investigações Filosóficas é um trabalho completo e que os escritos de Wittgenstein de 1946 em diante representam em certa medida partidas para novas direções',* **uma opinião que eu concordo**. Se Wittgentein teria ou não incorporado a Parte II no texto da Parte I, o fato é que ele não o fez. A Parte II não é parte do mesmo livro (HACKER, 2000, p. xvi-xvii, **grifo-nosso**).

A influência da leitura de von Wright em Hacker acabará influenciando, por consequência, a 4ª edição das *PU.* Nela, Hacker, agora junto de Joachim Schulte, continua seguindo um raciocínio muito próximo aquele afirmado em 1982 por von Wright:

> Quaisquer que sejam as intenções finais que Wittgenstein tinha, o fato é que o mais próximo que ele conseguiu chegar de completar as *Investigações Filosóficas* foi no texto presente que consiste dos §§1-693. Isto é, nos cremos, aquilo que deveria ser conhecido como as *Investigações Filosóficas* de Wittgenstein (HACKER; SCHULTE, 2009, xxiii).

A posição dos editores em relação à descontinuidade dos escritos pós-1945 com as *PU* não é carente de pares no meio acadêmico. Embora com objetivos distintos, podemos citar aqui o trabalho de Danièle Moyal-Sharrock, que alguns anos antes seguiu a mesma tendência em considerações sobre a Parte II. "Eu concordo com Peter Hacker que aquilo que foi publicado como Parte II das *Investigações Filosóficas* não deveria ter sido incorporado naquele trabalho (MOYAL-SHARROCK, 2004a, p. 207, nota 2)."

Além de considerar que a Parte II não deveria ser incluída nas *PU,* Moyal-Sharrock radicaliza a leitura de von Wright estabelecida em 1982. Ela chega a argumentar que o pensamento de Wittgenstein toma novos rumos essencialmente distintos após a versão final das *PU*, dando origem aquilo que ela chama de *terceiro Wittgenstein* (ou seja, uma 'terceira fase' do pensamento que essencialmente se diferencia tanto do *Tractatus* quanto das *IF*).

> [...] A visão mais geral da filosofia de Wittgenstein é aquela que a divide em duas fases distintas. [...] Esta divisão não reconhece suficientemente que *após* o trabalho este seminal [*PU*], Wittgenstein toma novos fundamentos. Embora nenhuma retratação esteja em

> questão aqui, acredito que o desenvolvimento do pensamento de Wittgenstein é suficiente para garantir a distinção de uma fase pós-*Investigações*, um *terceiro* Wittgenstein (MOYAL-SHARROCK, 2004b, p. 1).

Diferente do que afirmaram Hacker, Schulte e Moyal-Sharrock, alguns personagens da literatura especializada de Wittgenstein – tal como Geach (1988) e Venturinha (2007) – argumentam que os escritos sobre psicologia redigidos após 1945 não constituem uma 'nova temática' ou um 'trabalho paralelo' da filosofia de Wittgenstein, mas que eles são um *desdobramento* das reflexões iniciadas nas *PU*. Noutras palavras, com base no trabalho desses comentadores, iremos expor agora a tese que afirma que Wittgenstein trabalhou até o fim de sua vida no texto das *PU* e, consequentemente, que os escritos sobre filosofia da psicologia (incluindo a antiga 'Parte II') são parte desse contínuo trabalho, não novos eixos temáticos.

Como dissemos anteriormente, um personagem que serve de base inicial a favor da tese de continuidade das *PU* para com os escritos sobre filosofia da psicologia é Peter Geach, que fez anotações esclarecedoras sobre as últimas aulas ministradas pelo filósofo nos anos de 1946-7.[12]

Geach defendeu abertamente a ideia de que Wittgenstein, caso não fosse pego por sua morte prematura, teria revisado até o fim de sua vida aquilo que é hoje conhecido 'versão final' das *PU*.[13] Sobre isso, ele pondera:

> Nos últimos anos de sua vida ele [Wittgenstein] estava trabalhando de forma árdua nas *Investigações Filosóficas*. [...] A "Parte I" das Investigações estava completa quando Wittgenstein morreu, e nós já tínhamos visto o MS daquilo que agora é impresso como "Parte II"; Wittgenstein pretendia revisar as páginas finais da Parte I para incorporar o novo material, mas morreu antes de realizar tal revisão (GEACH, 1988, prefácio - xiii).

Além de Geach, outro fator corrobora – e muito – com a ideia de que os escritos de psicologia seriam escritos, ao menos em certa medida, como parte da revisão das *PU*: a menção direta ao livro ou a parágrafos das *PU* nos textos pós-1945. Em seu artigo "Against a third Wittgenstein", Nuno Venturinha (2007) destaca ao menos três passagens presentes nos MS 137 e MS 138, escritos entre 1948 e 1949, corroboram com a ideia de que Wittgenstein teria pensado o material da psicologia como um desdobramento das *PU*.[14]

A primeira passagem, datada de 9 de novembro de 1948, se encontra no §150 dos *Últimos Escritos sobre Filosofia da Psicologia, volume I* (LWPPI, vide MS 137, p. 32b). Nela, Wittgenstein afirma: "não é casual que eu empregue **neste livro**

[12] As anotações das aulas dos anos 1946-7 de Wittgenstein feitas por Peter Geach, Kanti Shah e A.C. Jackson foram publicadas em 1988, sob o título *Wittgenstein's Lectures on Philosophy of Psychology – 1946-1947*.

[13] Vale lembrar que Peter Geach era marido de Elizabeth Anscombe, editora da primeira versão das IF junto com Rhees. Sempre é bom relembrar que, além de Geach, os próprios Anscombe e Rhees também sustentam a tese de que a Parte II teria sido incluída na Parte I caso Wittgenstein não tivesse falecido precocemente.

[14] Tais manuscritos foram publicados em forma de livro e receberam o título de *Últimos Escritos sobre a Filosofia da Psicologia, volume 1.* (LWPPI).

tantas proposições interrogativas" (**grifo-nosso**). É bastante razoável pensar que 'neste livro' seja uma referência direta às *Investigações Filosóficas*. Porém, alguém poderia questionar: mas não estaria Wittgenstein referindo-se a outro livro, um livro posterior às *IF,* por exemplo? Para sanarmos essa dúvida, tomemos uma anotação do autor feita alguns dias depois - 28 de novembro - e que está presente no §340: "Se o jogo de linguagem, a atividade, o de construir uma casa, por exemplo (como no nº 2), fixa o emprego de uma palavra, o conceito de emprego é elástico relativamente à atividade (WITTGENSTEIN, 1982, §340)."

A referência ao 'nº2' presente na passagem só pode ter sentido se a tomarmos como sendo relacionada ao §2 das *PU*. Afinal, é nele que encontramos o clássico exemplo da 'linguagem dos construtores'. A referência à linguagem dos construtores é um forte indício de que os parágrafos iniciais das *PU* continuavam presentes no norte especulativo das reflexões sobre filosofia da psicologia pós-1945.

Além da referência à linguagem dos construtores presente no §340, temos também outra referência presente no §833, datada de 7 de fevereiro de 1949 (vide MS 138, p. 16a), na qual Wittgenstein faz menção ao 'jogo de linguagem de 8'.

> Mas o que significa 'convencer-se de algo?' Para o percebermos, temos de proceder a jogos de linguagem simples com esta palavra. – Como se convence alguém, no **jogo de linguagem 8**, de que ali ficam tantas e tantas lajes? Como nos convencemos de que 6+6=12? Etc. (WITTGENSTEIN, 1982, §833, **grifo-nosso**).

Ao tomarmos o termo 'jogo de linguagem 8", juntamente com outros termos que também estão presentes no parágrafo, como "jogos de linguagem simples" e "lajes", não nos resta dúvida sobre a relação direta com o §8 das *PU*.[15] Uma vez que a menção ao §8, assim como a menção ao §2, é de relevância central para o entendimento das anotações nesses escritos, é difícil crer que Wittgenstein não tivesse em mente um desdobramento das reflexões presentes no início das *PU.*

Os parágrafos incluídos nos *LWPPI* que fazem referência às *PU* servem como fundamento maior para a tese de Venturinha que afirma que Wittgenstein manteve-se ativo até sua morte com as questões iniciadas na primeira parte das *PU*. Essas passagens podem nos levar a concluir que parece pouco provável que os escritos sobre a filosofia da psicologia redigidos após 1945 possam ser lidos de forma 'temática', como algo separado ou pertencendo a livros distintos. Também poderia-se colocar em xeque a ideia de que as *PU* fossem um livro concluído por Wittgenstein naquilo que ficou conhecido como 'versão final' (TS 227), em

[15] Consideremos uma extensão da linguagem 2. Fora as quatro palavras 'cubos', 'colunas', etc., conteria uma série de palavras que seria empregada como o negociante no §1 emprega os numerais (pode ser a série das letras do alfabeto); além disso, duas palavras, que podem ser 'ali' e 'isto' (porque isto já indica mais ou menos sua finalidade), e que são usadas em combinação com um movimento indicativo da mão; e finalmente um número de modelos de cores. A dá uma ordem de espécie: 'd-lajota-ali'. Ao mesmo tempo faz com que o auxiliar veja um modelo de cor, e, pela palavra 'ali', indica um lugar de construção. Da provisão de lajotas, B toma uma da cor do modelo para cada letra do alfabeto até 'd' e a leva ao lugar que A designa. – Noutra ocasião, A dá a ordem: 'isto-ali'. Dizendo 'isto', aponta uma pedra. Etc (WITTGENSTEIN, 1982, §8)

1945/46. Porém, aceitar a referência a passagens anteriores como fundamento para uma "tese de continuidade" nos traz graves problemas teóricos.

Se levarmos em consideração o argumento das referências diretas de Wittgenstein aos seus jogos de linguagem (2) e (8), sustentado por Venturinha, como prova de que ele ainda estava trabalhando nas *IF*, e que, portanto, todo o material pós-1945 é também parte integrante das *IF*, e não uma nova frente de trabalho diferente desta, o mesmo raciocínio também tem que valer para outros textos de Wittgenstein, como por exemplo, o texto presente nas atuais *Observações sobre os Fundamentos da Matemática* (OFM). Afinal, tal texto foi escrito entre 1938 e 1944, no qual há várias referências diretas ao "jogo de linguagem (2)", das *IF*.[16] Mais ainda: se, então, todos os textos escritos entre 1936 e 1951 (porque também há referências ao "jogo de linguagem (2)" no texto do "*Sobre a Certeza*", são parte integrante das *IF*, então todo o *Nachlass* (1929-1951) poder entendido como uma contínua revisão das *IF*, pois este é seguramente composto com todo esse material literário, pelo que podemos retraçar de suas várias camadas textuais. Porém, o fato de um trabalho continuar a se engajar com as mesmas questões e assuntos como outro trabalho (anterior), não implica que o trabalho não possa, ao mesmo tempo, partir em novas direções e ser um trabalho próprio. Isso se aplica, por exemplo, a versão "*Urfassung*" das *Investigações Filosóficas*, que pode ser entendida como uma continuação do *Livro Marrom*, enquanto ainda é também um trabalho separado e parte em novas direções muito diferentes. A conclusão de que é parte do mesmo trabalho a ser publicado permaneceria injustificada, mesmo que o trabalho não partisse para uma nova direção.

Além da generalização apressada baseada na menção a trabalhos anteriores, o fato de que Wittgenstein, ao produzir a Parte II e outros escritos, continuou a trabalhar em paralelo nos textos da Parte I – e até mesmo se refere explicitamente a eles – não implica que os dois estejam juntos. Uma explicação igualmente plausível seria dizer que Wittgenstein pode ter desejado continuar a se envolver com os pensamentos da Parte I, seja como parte apenas de uma revisão dos textos da Parte I, seja por já estar trabalhando em algo novo e diferente disso. Não há contradição entre fazer as duas coisas ao mesmo tempo, todos trabalhamos em diferentes artigos e livros ao mesmo tempo.

Por fim, o Venturinha (2007) reutiliza justamente o ponto frequentemente mencionado na literatura secundária de que Wittgenstein estava parcialmente descontente com as últimas 150-200 seções da Parte I e invoca, nesse contexto, a opinião de Rhees e Anscombe de que a Parte II deveria substituir estas seções. Mas, embora Venturinha mostre que Wittgenstein no final dos anos 40 frequentemente se refira e discuta seções da Parte I, ele não mostra que ele discute precisamente as últimas 150-200 seções e seus tópicos - o que nós esperaríamos se fossem eles os quais a Parte II deveria substituir.

Além disso, o trabalho de Venturinha não nos dá uma visão mais detalhada de como o texto da Parte II está relacionado com a Parte I, de modo que a

[16] Por exemplo, nas *Observações sobre os Fundamentos da Matemática* (WITTGENSTEIN, 1956), Parte III, § 80; Parte VI, § 40; e Parte VII, § 71.

afirmação de que ele se encaixa e substitui filosoficamente parte dele permanece obscura e não sustentada. Também não apresenta nenhuma evidência textual que mostre que Wittgenstein no TS-234 ou manuscritos próximos estivessem realmente trabalhando na substituição/complementação da Parte I. Essa linha de argumentação deveria estar em vigor caso se queira mostrar que a Parte II pertença ou mantenha relação à Parte I. Afinal, apesar do caráter fragmentário dos textos pós-1945, é indiscutível a relevância dos pensamentos expostos nesses escritos desse período para um olhar mais fiel daquilo que Wittgenstein poderia ter em mente no que diz respeito à conclusão das *PU*, além dos caminhos que seguiria em sua investigação caso tivesse uma vida mais longa.

Saldo do percurso: algumas considerações

Esperamos que o caminho feito até aqui permita ao leitor chegar a algumas conclusões relacionadas às *Investigações Filosóficas* e aos escritos sobre filosofia da psicologia redigidos pós-1945. Isto, obviamente, inclui a antiga 'Parte II' e, consequentemente, a uma avaliação da decisão editorial presente na 4ª edição.

Podemos concluir que o texto final das *Investigações* é o resultado de um longo processo de trabalho. Metaforicamente, podemos dizer que ele foi construído 'por camadas' – em cada nova versão Wittgenstein promovia alterações e realizava algumas inclusões de novos materiais. Dos primeiros parágrafos da versão inicial (*Urfassung*) escritos entre 1936-37 até o parágrafo 693 da versão final (*Spätfassung*), ditados no ano letivo de 1945-46 lá se vão quase 10 anos. Levando isso em consideração – e incluindo todas as mudanças feitas nesse período -, é difícil apontar para algo que nos faça concluir que Wittgenstein terminaria ali seu trabalho de revisão. É muito mais provável pensar que ele continuaria nessa tentativa, embora ele mesmo tivesse se desiludido da tarefa de terminar essa revisão em vida.

No que diz respeito à inclusão do TS 234 nas *PU*, temos o seguinte: seja ele chamado de Parte II (tal como a nomearam Anscombe e Rhees), seja de *Filosofia da Psicologia – Um fragmento* (como propõe Hacker e Schulte), parece claro que estamos aqui diante de um material que não foi editado com a mesma propriedade, se assim podemos dizer, do que os §§1-693. Todavia, a menção a parágrafos das *Investigações* feitas em manuscritos pós-1945 pode nos remeter a ideia de uma continuidade das reflexões de Wittgenstein presentes nas *Investigações Filosóficas*, o que aparentemente advogaria contra a ideia de uma "partida para nova direção" (como afirma Wright) e a decisão de Hacker e Schulte de separar o TS-234 das *Investigações.*

Entretanto, ainda que a ideia de uma 'terceira fase' do pensamento de Wittgenstein nos soe bastante forçoso, visto que de fato Wittgenstein se dedica a tópicos relacionados ao mesmo eixo no TS-234, a saber, o problema do significado, o que temos é que a mera menção a passagem das *Investigações* nos escritos pós-1945 não constitui, tal como acredita Venturinha (2007), um argumento suficiente para a ideia de uma *continuidade* de abordagem de Wittgenstein para com os mesmos problemas presentes nas *Investigações.*

Nesse sentido, tanto a decisão de Hacker e Schulte (2008) de separar o TS-234 das *Investigações* sob a justificativa de serem trabalhos rigidamente distintos, quanto a leitura continuísta de Venturinha, que pressupõe uma intima conexão dos temas ali apresentados com aqueles presentes nas *Investigações*, possuem incoerências teóricas e limitações documentais. Esses entraves e percalços apenas evidenciam que a querela interpretativa sobre o *locus* do TS 234 no pensamento de Wittgenstein continua sendo, décadas após a primeira edição das *Investigações,* um tópico ainda atual e polêmico nos estudos sobre o filósofo.

Referências

BIGGS, Micheal; PICHLER, Alois. *Wittgenstein: two source catalogues and a bibliography*. Catalogues of the Published Texts and of the Published Diagrams, each Related to its Sources. Working Papers from the Wittgenstein Achieves at the University of Bergen, n.7, 1993.

ERBACHER, Cristian. Editorial approaches to Wittgenstein's Nachlass: towards a historical approach. *Philosophical Investigations*, n. 38, p. 165-198, 2015.

GEACH, Peter; SHAH, Kanti; JACKSON, A.C. *Wittgenstein's Lectures on Philosophy of Psychology 1946-1947*. Harvester-Wheatsheaf: Hertfordshire, 1998.

HACKER, P. M. S.; SHULTE, Joachim. The text of the Philosophische Untersuchungen. In: WITTGENSTEIN, Ludwig. *Philosophical Investigations*. Trad. G. E. M. Anscombe, P. M. S. Hacker e Joachim Schulte. 4ªEd. Oxford: Ed. John Willey & Sons [Blackwell Publishing], 2009.

HACKER, P. M. S. *Wittgenstein: Mind and Will*, Part II: Exegesis. Oxford: Blackwell, 2000.

_____. *Wittgenstein: Comparisons and Context*. Oxford: Oxford University Press, 2013.

MALCOLM, Norman. *Ludwig Wittgenstein - a memoir*. 2nd Edition. Oxford University Press: Oxford, 1984.

MOYAL-SHARROCK, Danièle. *Understanding Wittgenstein's On Certainty*. New York: Palgrave MacMillan, 2004a.

_____. Danièle. *The third Wittgenstein*: the post-investigation works. New York: Ashgate Publishing, 2004b.

VENTURINHA, Nuno. Against a third Wittgenstein. In: HRACHOVEC, H.; PICHLER, A.; WANG, J. *Papers of the 30th IWS*. ALWS Archives: A selection of papers from the International Wittgenstein Symposia in Kirchberg am Wechsel, 2007.

WITTGENSTEIN, Ludwig. *Wittgenstein's Nachlass*. The Bergen Electronic Edition. Bergen: OUP, 2000.

WITTGENSTEIN, Ludwig. *Bemerkungen Über Die Grundlagen Der Mathematik - Remarks on the Foundations of Mathematics*. Oxford: Basil Blackwell, 1956.

WITTGENSTEIN, Ludwig. *Philosophische Untersuchungen: Kritisch-genetisch Edition*. Notas de SCHULTE, Joachim; NYMAN, Heikki; von SAVIGNY, Eike; von WRIGHT, G. H. Ed. Surkhamp, 2001

_____. *Philosophical Investigations*. Trad. G. E. M. Anscombe, P. M. S. Hacker e Joachim Schulte. 4ªEd. Oxford: Ed. John Willey & Sons [Blackwell Publishing], 2009.

_____. *Letzte Schriften über die Philosophie der Psychologie. Vorstudien zum zweiten Teil der Philosophischen Untersuchungen / Last Writings on the Philosophy of Psychology*. Vol. I. Preliminary Studies for Parte II of Philosophy of Psychology. von WRIGHT, G.H.; NYMAN, H. (ed.). Trad. C. G. Luckhardt e M. A. E. Aue. Oxford; Basil Blackwell Publisher Limited, 1982.

_____. *Letzte Schriften über die Philosophie der Psychologie - Band II – Das Innere um das Äußere 1949-1951 / Last Writings on the Philosophy of Psychology*. vol. II. The Inner and the Outer 1949-1951. von WRIGHT, G.H.; NYMAN, H. (ed.). Trad. C. Luckhardt e M. Auer. Oxford; Basil Blackwell Publisher Limited, 1992.

_____. Philosophie der Psychologie - Ein Fragment / Philosophy of Psychology – A Fragment. In: *Philosophical Investigations*. Trad. G. E. M. Anscombe, P. M. S. Hacker e Joachim Schulte. 4 ed. Oxford: Ed. John Willey & Sons [Blackwell Publishing], 2009.

von WRIGHT, Georg Henrik. The Origin and Composition of the *Philosophical Investigations*. In: von WRIGHT, G.H. *Wittgenstein*. Oxford: Blackwell, 1982.

_____. The Troubled History of Part II of the *Investigations*. In: SCHULTE, Joachim; SUNDHOLM, Göran (Eds.). *Criss-crossing a philosophical landscape*. Essays on Wittgensteinian Themes Dedicated to Brian McGuinness. Amsterdam: Editions Rodopi, 1992.

_____."Vorwort". In: WITTGESNTEIN, L. *Philosophische Untersuchungen*. Kritisch-genetische Edition. Suhrkamp, 2001.

Recebido em: 06-03-2019

Aceito para publicação em: 17-07-19

HOW TO MEASURE A QUALE

COMO MEDIR UM QUALE

OSVALDO PESSOA JR.[1]
Universidade de São Paulo (USP) – Brasil
opessoa@usp.br

ABSTRACT: According to the colored-brain thesis (or qualitative physicalism), sense data or qualia are real physical-chemical qualities, located inside the brain, possibly at a specific locus. Our hypothesis is that the seats of phenomenal consciousness have a structure and a materiality. According to the proposed view, a chromatic quale emerges when a certain pixel of the visual sensorium (the hypothetical subjective visual "screen", or Cartesian theater, with its specific materiality ω) is fed with a certain pattern Σ of spikes; a change in this pattern quickly changes the color that is subjectively generated. How could one manage to measure chromatic qualia? In principle, with nanoscopical techniques, one could capture all the patterns that fall on the sensorium, and transmit the information to other media. But this does not capture the qualia. However, if the patterns are made to fall on a tissue of the same kind, typically inside another person's brain, this other person will have roughly the same subjective experience as the first person. The model is used to explore two different situations involving qualia inversion. The paper also explores Cartesian materialism, and the claim that phenomenal time and space are identical to a region of physical time and space.

KEYWORDS: Colored-brain thesis. Materiality. Qualia. Qualitative physicalism. Sensorium.

RESUMO: *De acordo com a tese do encéfalo colorido (ou fisicismo qualitativo), os dados dos sentidos ou qualia são qualidades fisicoquímicas reais, localizados no encéfalo, possivelmente em um local específico. Nossa hipótese é de que as sedes da consciência fenomênica têm uma estrutura e uma materialidade. De acordo com a visão proposta, um quale cromático emerge quando um certo píxel cromático do sensório visual (a hipotética "tela" visual subjetiva, ou teatro cartesiano, com sua materialidade específica ω) é alimentado com um certo padrão Σ de espículas; uma mudança neste padrão rapidamente altera a cor que é gerada subjetivamente. Como se poderia medir qualia cromáticos? Em princípio, com técnicas nanoscópicas, poder-se-ia capturar todos os padrões que caem no sensório, e transmitir a informação para outro meio. Mas isso não captura os qualia. Porém, se os padrões caírem em um tecido da mesma espécie, tipicamente dentro do encéfalo de outra pessoa, então esta outra pessoa teria aproximadamente a mesma experiência subjetiva que a primeira. Este modelo é usado para explorer duas situações diferentes envolvendo inversão de qualia. O artigo também explora o materialismo cartesiano, e a afirmação de que o tempo e o espaço fenomênicos são idênticos a uma região do tempo e espaço físicos.*

PALAVRAS-CHAVE: *Tese do encéfalo colorido. Materialidade. Qualia. Fisicismo qualitativo. Sensório.*

[1] Departamento de Filosofia (FFLCH) – Universidade de São Paulo (USP).

1 Colored-brain thesis

The *colored-brain thesis* (in Portuguese, "tese do encéfalo colorido", see PESSOA, 2017) is the name given by Leopold Stubenberg (1998, p. 169) to the view that phenomenal qualities, or qualia, are "properties of the brain". H.H. Price (1932, p. 127) referred to this thesis as the "hypothesis that sense data are cerebral". Price mentioned that the Hegelian philosopher F.H. Bradley had found it hard to believe that "when I smell a smell I am aware of the stinking state of my own nervous system", as the latter criticized Oxford philosopher Thomas Case, who had advocated the view. Case (1888, p. 33) characterized sense perception as the "the immediate apprehension of an internal physical object inside the nervous system of a sentient being". Case's position, however, was not materialist or physicalist, since he considered that God created and governs the world, and that the internal sensible object, which he took to be physical, was different from the "internal operation" which apprehends it, which he took to be "psychical".

The psychologist Edwin Boring came close to the thesis in 1933, as he argued for the mind-brain identity thesis in the interwar context of sense-data theory. According to U.T. Place (2000, p. 1), "Boring moreover, was himself apparently committed to combining the identity theory with a phenomenalist account of sensory qualities which on Leibniz's principle of the Identity of Indiscernibles would commit him to the view that certain brain events are literally green, high pitched, warm, sour or putrid, which for a philosopher would constitute an immediate knockdown reductio ad absurdum of his position".

In summary, according to the colored-brain thesis, which might also be called "qualitative physicalism", sense data or qualia are *real physical-chemical qualities*, yet to be treated by physical theory. Qualitative physicalism encompasses three main theses: ontic physicalism, the reality of qualia, and a mind-brain identity thesis.

The subjective greenness we experience as we look at an avocado is not in the fruit, but in our brain. If that is so, why doesn't a neurosurgeon ever see a green patch in our brain? The avocado appears green because of the wavelengths of light that it reflects, but our brain tissue does not have the same reflectance spectrum, so it doesn't appear green to an external observer. Subjective color has nothing to do with light (except for the detailed causal connection between the two): our brains are dark.

2 What is physical?

There is no consensual definition of what "physical" is, but one feature to be encompassed by the term is that it involves processes in space, time, and scale (micro-macro). So if a chromatic quale is taken to be physical, it should be located inside the brain, possibly at a specific locus, and our phenomenal visual field should have a specific size, for example around the scale of 10 cm^2, which is roughly the size of the retina (KOLB, 2007). We will adopt this localizationist hypothesis, although a more holistic view is also tenable.

In the Mary's room thought experiment (JACKSON, 1982), it is assumed that inside her black-and-white room, Mary has complete "physical knowledge" about colors, that is, a complete linguistic-quantitative descriptive knowledge (which leaves out only knowledge by acquaintance of colors). The question is, does she know all there is to know about colors?

When Mary finally leaves the room and notices for the first time a patch of green pigment painted on a wall (acquaintance), does she acquire new knowledge? And then when someone tells her that the patch is green (knowledge by acquaintance), is there any new element added to her knowledge about green? The usual answer is yes. This indicates that there is a difference between "physical knowledge" (linguistic-quantitative description and experimental capacity of manipulation) of an element and the acquaintance with it. This difference is what is called *qualia*, or subjective qualities.

Jackson's "knowledge argument" thus leads to the thesis that *there is non-physical knowledge about the world*. We note that this thought-experiment defines "physical knowledge" in a specific way. But accepting this definition, one concludes that the knowledge of qualia is "non-physical knowledge". But could one also conclude that *qualia are non-physical entities*? Such a view may be associated with David Chalmers (1996, p. 162), who considers a quale a "property" (p. 359). To conclude that qualia are non-physical, one would have to add another hypothesis to the argument, that "if something is knowable and if it is physical, then it is physically knowable".[2] One infers from this that there is something non-physical that is knowable, which would be *qualia*. Therefore, physicalism would be false.

Accepting Jackson's argument, the resulting dilemma is either to admit that physicalism is false, or to regard qualia as physical (rejecting Chalmers' hypothesis). This second alternative results in qualitative physicalism. A similar point is stressed by Owen Flanagan (1993, p. 98), who considers that the Mary's room thought experiment refutes linguistic physicalism, but not metaphysical (or ontic) physicalism.

3 Structure and materiality

As an example, a computer model of a hurricane captures many relations between portions of water and wind within the hurricane, but the computer which realizes the modelling is not itself wet. Wetness is part of the "materiality" of the hurricane, involving real water. In the same token, we adopt the anti-functionalist ("psychosubstantialist") view that a computer model of the brain cannot have consciousness, because consciousness depends on a certain materiality that is present in biological tissues, but apparently not in silicon-based computers.

[2] Let us define the following predicates: Fx: "x is a physical thing"; Cx: "x is knowable"; Px: "x is physically knowable", where by definition: (1) $(\forall x)\,(Px \rightarrow Cx)$; (2) $(\forall x)\,(Px \rightarrow Fx)$. Consider the following propositions: (3) $(\exists x)\,(Cx \wedge \neg Px)$ (Jackson's knowledge argument); (4) $(\forall x)\,((Cx \wedge Fx) \rightarrow Px)$ (lemma leading to Chalmers' hypothesis). From (3) and (4), one infers $(\exists x)\,(\neg Fx)$.

Thus, our hypothesis is that the seats of phenomenal consciousness have a *structure* and a *materiality*. The organizational structure Σ is given by the spatial distribution of the neurons and supporting cells and molecules, and by their interaction, including the temporal spikes that flow through the neural networks. This organization is reproducible in a machine and may be theoretically represented (by Mary inside her room) and manipulated. The material component ω is specific to the biological tissue, and is probably lacking in silicon-based machines (which have their specific materiality, of course). This seems to be in agreement with Searle's (1992) "biological naturalism".

A view close to qualitative physicalism has been defended by phenomenalists such as Schlick (1918) and Russell (1927), who combine sense-data or sensorial elements with the view that physics only has access to the relations between things, a view sometimes called "physical structuralism" (or "structural realism"). Since, for Russell, the "percepts" are the only part of the physical world that we know in a non-abstract way, he concludes that the physical world is made of the same elements that we experience consciously. Schlick and Russell are not exactly materialists (which is the case of qualitative physicalism). Russell's views have had a recent impact in the philosophy of mind (see Alter & Nagasawa, 2015), generating a class of views classified as "Russellian monisms", which considers that there exist "qualities", "quiddities", "inscrutables" or "things-in-themselves" that permeate all of the physical world, and to which we don't have direct access, except for the sense data.

Qualitative physicalism maintains that a subjective sensation is *identical* to a real physical quality, probably of electrical/chemical nature, possibly localized in space, time and scale (see PESSOA, 2017). By extension, one may assume that matter is endowed with these qualities or quiddities, and that these combine in some complicated way in the brain tissue, generating the qualitative complex that we experience daily. This possibility is also considered by Herbert Feigl (1967), commenting on the "*pan-quality-ism*" proposed by Stephen Pepper.

Another tradition that is close to qualitative physicalism is what Chalmers (2015) called "panprotopsychism". This tradition is represented by the English physicist and philosopher William Clifford, who in 1878 defined what he called *mind-stuff*, present even at the atomic level, and the composition of which would generate conscious mind in human beings. This view had an influence on American neorealism, and the panprotopsychism of Durand Drake (1933) is very close to qualitative physicalism (see PESSOA, 2017).

4 The chromatic sensorium

The colored-brain thesis implies that the subjective visual field exists in physical space, either as a direct physical screen (topologically distributed just like we experience it) or as a convoluted image. The first (and simpler) hypothesis amounts to the postulation of an identity between phenomenal space and time, on the one hand, and a region of physical space and time, on the other (in accordance

with psychophysical isomorphism, see KÖHLER, 1943, pp. 61-2). According to this hypothesis, a chromatic quale emerges when a certain "pixel" of the subjective visual screen is fed with a certain pattern of spikes; a change in this pattern quickly changes the color that is subjectively (and physical-chemically) generated. Such patterns (which generate colors downstream) are probably produced in the visual cortical area V4. Only when such spatial-temporal patterns arrive at the hypothetical sensorium (the subjective arena, or Cartesian theater) are the physical-chemical qualities created, and somehow the core ego becomes aware of the scene (how this happens, of course, is still an open problem).

Such a screen may be either located in a small region (localizationism) or distributed over a large extent of the brain (holism). Following the first option, if one assumes that the density of information in the retina is comparable to that in the chromatic screen, such an internal screen might have a size of around 10 cm^2. The retina has an area of around 10.94 cm^2 (KOLB, 2007), generating a binocular visual field of around 200° × 135°, so the internal screen might occupy a region of approximately 4.0 cm × 2.7 cm. Another estimate can be made, of the size of the immediate correlate of conscious vision, by assuming that each subjective visual pixel is generated by a single biological cell. Considering the span of our binocular visual field of 200° horizontally, and taking the limit of angular resolution of the human eye as being roughly 1 minute of arc (1/60 of a degree), if this limit corresponds to a single cell of around 4 micra, one computes the horizontal size of the visual field as being around 4.8 cm, consistent with the previous estimate.

Such an image is most probably of an electrochemical nature, occurring in a special tissue (such as the reticular formation in nuclei of the thalamus) subject to spatial and temporal patterns of electrochemical spikes. In the case of subjective color sensations, Valberg (2001) and others have argued that the four (or six) basic colors, which constitute opposing colors in the sense of Hering (green-red, blue-yellow, plus black-white), are not implemented in a pure form in the retina or in the lateral geniculate nucleus. This indicates that they are generated further downstream, and one might speculate that this takes place in the immediate neural correlate of vision (the visual sensorium). Since greenness and redness (for instance) cannot be instantiated at the same time, in the same pixel, this might suggest that some electrochemical system of the cell is either in an electrically positive or negative state (associated to greenness or redness), or in a neutral state, and in only one of such states at a given time (see HERING, 1913, pp. 59-61). According to our model, what determines these states is the temporal pattern of spikes that come into the cell from different external dendrites. Quantities generate qualities in a modulated electrochemical process. This would be different from the opponent cell structure in the parvocellular layers of the lateral geniculate nucleus (in the thalamus), where "red-ON cells", "green-OFF cells", "red-OFF cells" and "green-ON cells" process the incoming information.

Above the generation of qualitative chromatic pixels, one must also try to account for higher levels of representation, which identify smaller and larger patterns, ranging from localized corners all the way up to the figure's Gestalt or

overall form. This might be implemented by layers of cells behind (or on both sides of) the layer of chromatic pixels.

In the philosophical literature, the chromatic screen (and analogous sensoria for other sense modalities) has been referred to as the "Cartesian theater" (Dennett, 1991, p. 107) or the "picture story" (PYLYSHYN, 2007, p. 121). The hypothesis of such a sensorium is criticized because, if there is a homunculus that is sitting in the audience and watching the picture show, how is one to explain what takes place inside the homunculus, without incurring an infinite regress? Of course one should not postulate a separate audience for the theater: somehow the core self must be *built upon* the Cartesian theaters (for the different modalities), breaking the separation between subject and object. The hypothesis presented here falls into what Dennett (1991, p. 207) has called "Cartesian materialism", the view "that there is a crucial finish line or boundary somewhere in the brain, marking a place where the order of arrival equals the order of 'presentation' in experience because what happens there is what you are conscious of".

5 Measuring qualia

Assuming that the model is correct, how could one manage to *measure* chromatic qualia? In principle, with nanoscopic techniques, one could capture all the patterns that fall on the tissue of the sensorium, and transmit the information to other media. But this does not capture the qualia. However, if the patterns are made to fall on a tissue of the same kind of materiality, typically inside another person's brain, this other person will have roughly the same subjective experience as the first person! This is how a subjective experience may be shared with other creatures of the same biological species, at least in a rough sort of way.

The notion of measurement, for qualia, is different from the usual measurements in physics, such as the ratio of two masses, which only capture the *relations* between entities (the aforementioned structural or organizational aspect of the world). As mentioned above, we consider that a quale is a "thing in itself" (involving a materiality). One cannot directly capture a quale and put the picture on the wall. But in principle it is easy enough to feed the patterns of one brain into another brain of similar material constitution. One could generate a Kodachrome photograph to be put on the wall, which causes in us a similar subjective experience as the one lived by another subject, but the photograph itself is devoid of chromatic qualia.

Ramachandran & Hirstein (1997, pp. 432-33) arrived at a similar conclusion, inspired by talks with Francis Crick. They imagined a cable or "bridge of neurons" connecting visual area V4 of a person with normal vision to area V4 of a color-blind person due to problems in the retina, and concluded that the blind person would be able to experience the same chromatic qualia as the first person. This, however, does not mean that area V4 is the "chromatic sensorium", since it could be only an intermediate stage in the causal chain leading to the experience of qualia. Let us explore this considering another thought-experiment.

6 Modelling qualia inversion

To illustrate our model that qualia are produced when a certain pattern of neural spikes Σ falls upon a certain region of the brain (the sensorium) of specific materiality ω, consider Dennett's (1988, pp. 49-50) "intuition pump" (thought experiment) called *the Brainstorm machine*, in which someone (call him Dan) has access to the brain of another person (say Chloe) who inverts the color spectrum.

> Suppose [...] there were some neuroscientific apparatus that fits on your head and feeds your visual experience into my brain [...]. With eyes closed I accurately report everything you are looking at, except that I marvel at how the sky is yellow, the grass red, and so forth. Would this not confirm, empirically, that our qualia were different? But suppose the technician then pulls the plug on the connecting cable, inverts it 180 degrees, and reinserts it in the socket. Now I report the sky is blue, the grass green, and so forth. Which is the 'right' orientation of the plug? Designing and building such a device would require that its 'fidelity' be tuned or calibrated by the normalization of the two subjects' reports – so we would be right back at our evidential starting point. The moral of this intuition pump is that no intersubjective comparison of qualia is possible, even with perfect technology. (DENNETT, 1988, pp. 49-50).

According to our model, at least two things can happen, as Chloe looks at the grass and has the subjective experience of red.

Case (I): assume the sensorium of both Chloe and Dan are the same: $\omega_C = \omega_D$, but that Chloe sees the grass red because the spike train generated in area V4 and fed into the sensorium is different from the spike train generated in Dan's brain, which leads to the green quale: Σ_C (grass) $\neq \Sigma_D$ (grass). Dan has access to Chloe's brain, but he can only receive her spike train Σ_C, not having access to the material substrate ω_C. As Σ_C falls on ω_D, Dan will have the same subjective color experience as Chloe, i.e. he will see the image of the grass as red, since both sensoria are identical (and supervenience guarantees that the identical brain states generate identical mental states).

Case (II): assume the spike trains generated for both of them are the same, $\Sigma_C = \Sigma_D$, when looking at grass, but that Chloe has a red experience because her material substrate (sensorium) is in a different chemical state from that of Dan, so that $\omega_C \neq \omega_D$. In this case, Dan will receive from Chloe the same spike train that he usually sees when looking at grass, so his experience of the grass Chloe is seeing will be green, i.e. different from Chloe's subjective experience, which is red. He will have no clue that she is subjectively inverting the spectrum.

Case (III): there are other variations in which $\omega_C \neq \omega_D$ and $\Sigma_C \neq \Sigma_D$, but we will ignore them.

The situation described by Dennett would have to correspond to case (I), since Dan marvels at the color changes while having access to Chloe's brain (in case II he notices no change). In this case, the qualia that both of them experience

when looking directly at grass are clearly different. But the quale Dan experiences *when looking inside Chloe's head* is the same as that which she is experiencing at the time (contrary to what happens in case II).

Now, the assumption that the technician switches the plug really doesn't apply to case (I), since we have assumed that both sensoria are identical, and because the spike train is a mechanical (geometrical) property of the world, the determination of which does not depend on previous calibration of the subjects' reports.

In case (II), however, as noted above, Dan would have no way of knowing what color Chloe is experiencing. Neuroscientists could infer that she might be having a different experience, by noticing chemical differences in her sensorium (in relation to Dan's). But exactly what qualia each is experiencing is something only theory could tell us (up to its degree of confirmation). A third party might conclude that she is experiencing the same quale as either Chloe or Dan, as long as their sensoria are in the same physical-chemical state.

7 Projections in time and space

How is it that we project the visual scenery of our environment as something much larger than our bodies, given that the actual size of the representation we experience is of the order of only 10 cm^2?

To guide our reasoning, let us start with the notion of the *antedating of time*, which is connected to the definition of what is "now". According to Libet et al. (1979), it takes a small "period of adequacy" (between a fifth of a second and half a second) for a perceptual stimulus (such as a tactile pinch in the hand) to become conscious. Besides that, they showed that we "antedate" the occurrence of the perceptual event back to the past (by an amount equal to the period of adequacy), so as to correct our subjective ordering of events. Direct stimulus to the cortex is not antedated, so it was in relation to this that the mechanism of antedating was discovered. The hypothesis suggested by the authors was that a "marker" for the perceptual event was registered at the thalamic level at the time the stimulus first arrived there, so that the relative orderings of different perceptual events could be consciously evaluated in a correct manner after the period of adequacy.

We could say that this is one more example of how natural selection fine-tunes "the subjective construction of the representation of the world" so as to fit to the actual world. Since percepts of different intensities might have different periods of adequacy, such a mechanism guarantees in a simple way that the marks of the temporal ordering of the stimuli follow their times of arrival at the thalamus (before proceeding to conscious processing).

From a different perspective, this situation brings in the question of how to define the "present" time. To define it as "now" is circular. An ostensive definition such as "the time that I am saying the word 'now'" (or seeing a flea leap) requires one to mutter a word (or consciously process the observation of the leap), so that

the conscious perception of this act is delayed by the period of adequacy. Thus, the present cannot be captured simply by conscious observation. One has to introduce theoretical considerations, so that "the present associated with the flea's leap" would have to be "half a second before the time of conscious awareness of the leap", or "the time Libet's perceptual mark is registered in the thalamus", or a comparison with some other event.

The antedating of time is an instance of a "projection", i.e. the association of an object to a spatio-temporal setting established in relation to our body. What would be an analogous projection for the perception of space? Libet et al. (1979, p. 221) explore the analogy, but focus on the recognition that "the spatial form of a subjective sensory experience need not be identical with the spatial pattern of the activated cerebral neuronal system that gives rise to this experience". This issue bears on Köhler's aforementioned psychophysical isomorphism, and its alleged refutation in experiments on the neocortex, but it is still an open problem, given that the hypothetical visual sensorium has not yet been identified (and is probably not in the primary visual cortex).

The analogy with temporal antedating that we would like to stress is the subjective projection of our small "chromatic screen" to a size much larger than our own bodies. In other words, we experience our visual field as being outside our bodies, which is consistent with our actions in the world, but in fact (according to qualitative physicalism) it occupies a small region in our brains. Pathologies in this ability for projection seem to lead to certain forms of the Alice-in-Wonderland syndrome.

Besides the subjective construction of visual space directly from a small spatial patch in our brains, one might also mention the spatial representation that organizes our bodily senses, including the senses of touch, pressure, pain, temperature, vibration, and proprioception. The pathways arising from different parts of the body are spatially organized in the brain in a form that resembles the body, in regions called "somatosensory homunculi", to be found in the cortex (*parietal lobe*), thalamus (*ventralis posterior nucleus*) and cerebellum. In fact, there is a continuous somatotopical tissue organization all the way from the brainstem to the other regions of the brain, along the ascending lemniscal and anterolateral pathways. The seat of bodily consciousness (the somatosensory sensorium) could be located anywhere in these pathways, and once again the thalamus seems a good candidate, as stressed by a few authors, such as Bogen (1995) and Ward (2011).

An object of the sense of sound may be projected on external space, but the "tonotopic" organization of information which gives rise to subjective pitch is done roughly in one mirror-symmetric spatial dimension in Heschl's gyrus of the auditory cortex (DA COSTA et al., 2011).

As for the sense of smell, the thalamus does not seem to be a good candidate for the seat of conscious awareness. Around 400 olfactory receptors generate patterns in the glomerular layer of the olfactory bulb, which is further processed in the three-layered olfactory cortex. After this, there is "only a small contingent of

fibers going to the mediodorsal thalamus" (SHEPHERD, 2005, p. 166), with most of the pathway leading directly to the prefrontal cortex. There is no subjective projection of odors onto the "somatosensory-visual" external space, but Kumar et al. (2015) have drawn an "odor network", which represents subjective similarities between classes of odors.

Conclusion

The metaphysical solution to the mind-brain problem provided by qualitative physicalism (the colored-brain thesis) interprets qualia as real physical entities. Far from settling the problems in the field, the approach has given rise to a heuristics of investigation, based on the postulated identity between "phenomenal time and space" and "a region of physical time and space", which is typical of Cartesian materialism. A fundamental distinction is made between the organizational structure Σ of a physical process and its materiality ω. We have postulated that qualia arise from a material cellular substrate, when specific spatio-temporal patterns of electrochemical spikes are fed to the substrate. The issue of how to measure a quale is solved in a direct manner, by recognizing that the measurement of the pattern of spikes can generate a similar quale if it is fed to another identical material substrate, i.e. another human brain. Our model was applied to the Brainstorm-machine thought experiment, involving qualia inversion, yielding two distinct and interesting results. While neuroscience has not yet solved the issue of the immediate neural correlates of consciousness, we have explored some possibilities from a localizationist perspective.[3]

References

ALTER, Torin & NAGASAWA, Yujin (eds.). *Consciousness in the physical world: perspectives on Russellian monism*. Oxford: Oxford University Press, 2015.

BOGEN, Joseph E. On the neurophysiology of consciousness. I. An overview. II. Constraining the semantic problem. *Consciousness and Cognition*, v. 4, p. 52-62, 137-58, 1995.

BORING, Edwin G. *The physical dimensions of consciousness*. New York: Century, 1933.

CASE, Thomas. *Physical realism: being an analytical philosophy from the physical objects of science to the physical data of sense*. London: Longmans, Green & Co., 1888.

[3] This paper is based on the talk "Como medir um quale?", given at the *XI International Brazilian Meeting on Cognitive Science* (EBICC), in October 30, 2017, and on the lecture "Explorando o materialismo cartesiano", given by telecommunication at the *Encontro Cognição & Linguagem*, in November 6, 2018, both at the University of São Paulo. Much of the research stems from a grant from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – Bolsa de Produtividade). I am also grateful for the welcome given by the Dept. of History & Philosophy of Science and Medicine, Indiana University, where this version of the text was completed, and also for the useful comments of the referees.

CHALMERS, David J. *The conscious mind*. New York: Oxford University Press, 1996.

_____. Panpsychism and panprotopsychism. The Amherst lecture in philosophy, pp. 1-35, 2013. Available at: http://www.amherstlecture.org/chalmers2013/chalmers2013_ALP.pdf. Reprinted in ALTER & NAGASAWA (2015), pp. 246-76.

DA COSTA, Sandra; VAN DER ZWAAG, Wietske; MARQUES, José P.; FRACKOWIAK, Richard S. J.; CLARKE, Stephanie & SAENZ, Melissa. Human primary auditory cortex follows the shape of Heschl's gyrus. *Journal of Neuroscience*, v. 31, p. 14067-14075, 2011.

DENNETT, Daniel. Quining qualia. In: MARCEL, Anthony J. & BISIACH, E. (eds.). *Consciousness in contemporary science*. Oxford: Clarendon, 1988. pp. 42-77.

_____. *Consciousness explained*. New York: Little, Brown & Co., 1991.

DRAKE, Durand. *Invitation to philosophy*. Boston: Houghton Mifflin, 1933.

FEIGL, Herbert. Russell and Schlick: a remarkable agreement on a monistic solution of the mind-body problem. In: Roberts, George W. (ed.). *Bertrand Russell memorial volume*. Oxfordshire: Routledge, 1975. pp. 321-38.

FLANAGAN, Owen J. *Consciousness reconsidered*. Cambridge (MA): MIT Press, 1993.

HERING, Ewald. On the theory of nerve-activity. In: *Memory: lectures on the specific energies of the nervous system*. 4th ed., enlarged. Lecture given in May, 1898, in Leipzig. Chicago: Open Court, 1913. pp. 43-70.

JACKSON, Frank. Epiphenomenal qualia. *Philosophical Quarterly*, v. 32, p. 127-36, 1982.

KÖHLER, Wolfgang. *Gestalt psychology*. 2nd ed. New York: Liveright, 1943 [First edition: 1929].

KOLB, Helga. Facts and figures concerning the human retina, 2007. Available at: https://www.ncbi.nlm.nih.gov/books/NBK11556/.

KUMAR, Ritesh; KAUR, Rishemjit; AUFFARTH, Benjamin & Bhondekar, Amol P. Understanding the odour spaces: a step towards solving olfactory stimulus-percept problem. *PLOS One*, v. 10, n. 10, paper e0141263, 2015.

LIBET, Benjamin; WRIGHT, Elwood W., Jr.; FEINSTEIN, Bertram & PEARL, Denies K. Subjective referral of the timing for a conscious sensory experience: a functional role for the somatosensory specific projection system in man. *Brain*, v. 102, p. 191-222, 1979.

PESSOA Jr., Osvaldo. A tese do encéfalo colorido. Forthcoming in *Estudos Filosóficos* (São João del-Rei), 2017.

PLACE, Ullin T. Identity theories, 2000. Unfinished manuscript, edited by S. Schneider, available at the website *A field guide to the philosophy of mind*: http://host.uniroma3.it/progetti/kant/field/mbit.htm.

PRICE, Henry H. *Perception*. London: Methuen, 1932.

PYLYSHYN, Zenon W. *Things and places: how the mind connects with the world.* Cambridge: MIT Press, 2007.

RAMACHANDRAN, Vilayanur S. & HIRSTEIN, William. Three laws of qualia: what neurology tells us about the biological functions of consciousness, qualia and the self. *Journal of Consciousness Studies*, v. 4, p. 429-58, 1997.

RUSSELL, Bertrand. *The analysis of matter*. New York: Harcourt Brace & Co., 1927.

SCHLICK, Moritz. *General theory of knowledge*. Translation by A.E. Blumberg of the 2nd German edition of 1925 (1st ed. 1918). New York: Springer, 1974.

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

SHEPHERD, Gordon M. Perception without a thalamus: how does olfaction do it? *Neuron*, v. 46, p. 167-9, 2005.

STUBENBERG, Leopold. *Consciousness and qualia*. Amsterdam: John Benjamins, 1998.

VALBERG, Arne. Unique hues: an old problem for a new generation. *Vision Research*, v. 41, p. 1645-57, 2001.

WARD, Lawrence M. The thalamic dynamic core theory of consciousness. *Consciousness and Cognition*, v. 20, p. 464-86, 2011.

Recebido em: 06-03-2019

Aceito para publicação em: 25-06-19

A CONSCIÊNCIA COMO UMA PERCEPÇÃO DO MENTAL E O ESTATUTO DOS FENÔMENOS MENTAIS INCONSCIENTES NA PERSPECTIVA DE DAVID ARMSTRONG[1]

CONSCIOUSNESS AS A PERCEPTION OF THE MENTAL AND THE STATUS OF UNCONSCIOUS MENTAL PHENOMENA IN DAVID ARMSTRONG'S VIEW

TÁRIK DE ATHAYDE PRATA[2]
Universidade Federal de Pernambuco (UFPE) – Brasil
tarik.de_athayde_prata@alumni.uni-heidelberg.de

RESUMO: O artigo examina a teoria de David Armstrong sobre a consciência e sua concepção do inconsciente. Após uma discussão do caráter anti-cartesiano dessa teoria (seção 1), são discutidas as noções de consciência mínima e consciência perceptiva (seção 2), bem como o conceito de consciência introspectiva, que é o mais importante para Armstrong (seção 3). A conclusão é que, apesar do valor explicativo dos seus conceitos de consciência, Armstrong defende uma perspectiva insatisfatória a respeito do inconsciente, pois essa perspectiva não dá conta da real influência do inconsciente em nossa vida mental (seção 4).

PALAVRAS-CHAVE: Consciência. Inconsciente. Percepção. Introspecção.

ABSTRACT: *This paper examines David Armstrong's theory of consciousness and his view of the unconscious. After a discussion of the anti-Cartesian character of this theory (section 1), the concepts of minimal consciousness and perceptive consciousness are discussed (section 2), as well as the concept of introspective consciousness, which is the most important for Armstrong (section 3). The conclusion is that, despite the explanatory value of his concepts of consciousness, Armstrong holds an unsatisfactory view about the unconscious, since he does not account for its real influence on our mental lives (section 4).*

KEYWORDS: *Consciousness. Unconscious. Perception. Introspection.*

INTRODUÇÃO

Durante longo tempo, as diversas noções associadas à palavra "consciência" desfrutaram de um lugar de destaque nas concepções filosóficas (e científicas) sobre a mente. De acordo com o influente filósofo alemão Franz Brentano (1838-1917), ele mesmo um opositor da ideia de inconsciente, alguns milênios se

[1] O presente trabalho foi produzido no âmbito do projeto de pesquisa "Concepções teóricas da consciência e o problema dos fenômenos mentais inconscientes", registrado no Programa de Pós-Graduação em Filosofia da UFPE. Gostaria de agradecer aos pareceristas anônimos da revista **Sofia**, por suas importantes observações e sugestões.
[2] Departamento de Filosofia da Universidade Federal de Pernambuco (UFPE).

passaram na história da Filosofia antes que aparecesse um filósofo que defendesse a existência de fenômenos mentais inconscientes (cf. BRENTANO, 1924, p. 144; BRENTANO, 1995, p. 103),[3] e não se pode esquecer que René Descartes (1596-1650), o pensador de maior influência na reflexão filosófica sobre os fenômenos mentais,[4] caracterizava o pensamento (visto por ele como a essência da alma) como "tudo quanto está de tal modo em nós que somos imediatamente conscientes [*consci*]" (DESCARTES, 1979, p. 169 [AT, VII, p. 160; AT, IX, p. 124][5]).

Entretanto, por volta da primeira metade do século XX houve uma inversão nesse cenário, na medida em que a importância dos fenômenos inconscientes aumentou de maneira progressiva, e o principal responsável por essa mudança foi um aluno de Brentano (cf. SCHELLENBACHER, 2011; CATALDO-MARIA & WINOGRAD, 2013), o médico e criador da psicanálise Sigmund Freud (1856-1939). Como coloca John Searle, se as gerações passadas tendiam a ver a noção de consciência como não problemática, e a noção de inconsciente como misteriosa, ou até autocontraditória, nós invertemos os papéis, pois "depois de Freud, nós invocamos rotineiramente fenômenos mentais inconscientes para explicar os seres humanos, e achamos a noção de consciência misteriosa, e talvez mesmo anticientífica" (SEARLE, 1992, p. 151; SEARLE, 1997, p. 218). De fato, Freud insurgiu-se contra a identificação do domínio psíquico com a consciência, e defendeu que os processos mentais são, fundamentalmente, inconscientes.[6]

Mas o fato de que a consciência, hoje em dia, não é mais vista de modo tão óbvio como o fundamento da mente não significa que ela não seja um importante tema de investigação (Não se pode esquecer, aliás, que o próprio Freud elaborou ideias a respeito da consciência – Cf. CAROPRESSO, 2009). Na tradição da filosofia analítica, por exemplo, uma das mais influentes na contemporaneidade,[7] a temática da consciência adquiriu um lugar central, após décadas de negligência, causada pela influência do positivismo lógico e de várias outras concepções fisicalistas sobre a mente (cf. SEARLE, 1992, pp. 02-05; SEARLE, 1997, p. 08-13). Dentro da tradição analítica, se constituíram várias correntes de pensamento a respeito da

[3] "Até muito recentemente na história da ciência e da filosofia, a vida mental era considerada inteiramente, ou principalmente, consciente em sua natureza" (BARGH & MORSELLA, 2008, p. 73). As citações em língua estrangeira foram traduzidas para o português pelo autor do presente trabalho.

[4] "Representação e consciência são consideradas ainda hoje – apesar de Wittgenstein – como as características decisivas do mental. Não se pode encontrar um único homem que tenha sido mais influente para esse tipo de caracterização do mental do que Descartes. Suas teorias sobre esses assuntos não são aceitas por quase ninguém e menos ainda são os que as compreendem. Entretanto, o modo como ele trata filosoficamente esses temas é até hoje inconscientemente imitado por muitos" (KEMMERLING, 2005, p. 09).

[5] Entre colchetes estão indicadas as referências ao volume VII (latim) e ao volume IX (francês) da edição das obras de Descartes organizada por Charles Adam e Paul Tannery (AT).

[6] "A primeira dessas declarações perturbadoras da psicanálise diz que os processos psíquicos são em si e para si inconscientes, e que conscientes são apenas atos isolados e partes do todo da vida psíquica. [...] [A psicanálise] não pode aceitar a identidade do consciente com o psíquico. Sua definição do psíquico diz que seriam processos do tipo do sentir, do pensar, do querer, e ela tem que defender a existência de pensamentos e vontades inconscientes." (FREUD, 1982, p. 47).

[7] Como coloca Habermas: "a filosofia analítica adquire, nas décadas após a segunda guerra mundial, a sua posição imperial, a qual continua se mantendo até hoje através de Quine e Davidson" (HABERMAS, 1990, p. 12). O texto de Habermas foi publicado, originalmente, em 1988.

consciência (cf. VAN GULICK, 2012) que, no meu modo de entender, oferecem preciosas contribuições para a devida compreensão desse tema.

O caso é que, se podemos pensar a consciência como uma entidade ou como uma faculdade (que podemos designar com o nome abstrato "consciência"), podemos também pensá-la como algum tipo de *propriedade*, que pode ser designada pelo predicado monádico "... é consciente", ou pelo predicado diádico "... é consciente de ..." (cf. GENNARO, 1996, p. 03). E se a consciência é pensada como uma propriedade pertencente a certos estados mentais (que são, portanto, estados mentais conscientes), torna-se importante explicar *em que* essa propriedade consiste[8] e *como* ela surge, ou seja, explicar como estados mentais se tornam estados conscientes.

Essa formulação do problema já sugere que a consciência está sendo concebida como uma propriedade *não essencial* dos estados mentais que, eventualmente, a possuem; ou seja, uma propriedade sem a qual os estados mentais podem continuar a existir (de forma inconsciente). Uma concepção como essa, ou seja, uma concepção da consciência como uma *propriedade extrínseca* de certos estados mentais, articula-se com a derrocada do cartesianismo no campo das investigações sobre a consciência, ao mesmo tempo em que mostra a consciência como um tema importante de investigação, pois a questão a respeito de como essa propriedade surge é extremamente relevante para a compreensão de nossa vida psíquica. Em outras palavras: do fato de a consciência não ser uma propriedade essencial para a *existência* dos fenômenos mentais, de maneira que ela não seja o fundamento da mente, não se segue que a consciência não seja um tema importante de investigação, já que ela desempenha um papel importante em nossa existência psicológica.

Uma visão desse tipo a respeito da consciência (ou seja, que a vê como uma propriedade *extrínseca*) é o que encontramos nas chamadas "teorias de ordem superior" (*higher-order theories*), isto é, teorias segundo as quais a consciência como característica de um fenômeno mental decorre de seu monitoramento por um outro fenômeno mental, que é *acerca* do primeiro (GENNARO, 2004, p. 01-02; VAN GULICK, 2004, p. 68; VAN GULICK, 2012, pp. 47-49). E um dos principais expoentes das teorias de ordem superior foi o filósofo australiano David Armstrong (1926-2014), um dos maiores nomes da filosofia analítica no século XX, e autor cujas reflexões sobre a consciência tiveram grande impacto na filosofia da mente.

Armstrong via a consciência de estados mentais (no sentido mais interessante da palavra "consciência") como decorrente de uma *percepção do mental*,[9] um ponto de vista que, como será discutido a seguir, se coaduna de maneira simples e direta com a visão de que existem fenômenos mentais

[8] Conforme será explicitado a seguir, de acordo com as teorias de ordem superior, um estado mental possui a propriedade de ser consciente quando *nós* estamos conscientes *dele* (de modo que se trata de uma propriedade *relacional*).

[9] A percepção (interna) de um fenômeno mental é uma representação *meta-psicológica* no sentido de que essa percepção (que é um fenômeno psicológico) é *a respeito de* um outro fenômeno psicológico (cf. GENNARO, 2004, p. 01). O prefixo "meta" indica que um fenômeno psíquico (a percepção interna) se dirige a outro fenômeno psíquico (seja ele uma percepção externa, uma crença, um desejo, etc.).

inconscientes (cf. ARMSTRONG, 1997, p. 724). De acordo com ele: "consciência não é mais do que a *ciência*[10] (percepção) dos estados mentais internos pela pessoa que possui esses estados mentais. Se é assim, então a consciência é simplesmente *um outro* estado mental, dirigido aos estados internos originais" (ARMSTRONG, 1968, p. 94 – grifo acrescentado). É justamente nesse sentido, de se tratar de um estado mental dirigido a *outro* estado mental, que essas teorias são ditas "de ordem superior". A "ordem superior" nesse contexto, *não tem* a ver com nenhum aspecto valorativo. Essa expressão significa, simplesmente, que temos um estado mental de "segunda ordem" dirigido a um estado mental dito de "nível básico" ou de "primeira ordem" (cf. ROSENTHAL, 2008b, p. 835).

E a analogia da consciência com a percepção se mostraria no fato de que a percepção nos fornece *conhecimentos* a respeito de seus objetos, assim como a consciência (na sua noção mais sofisticada) nos torna *cientes de* nossos próprios estados mentais. Como esclarece Armstrong: "pela percepção sensorial nos tornamos cientes das ocorrências físicas que estão tendo lugar em nosso ambiente e em nosso corpo. Pelo sentido interno nos tornamos cientes das ocorrências que estão tendo lugar em nossa própria mente." (ARMSTRONG, 1968, p. 95). Nesta última passagem citada, Armstrong está se referindo àquela que, na sua concepção, é a forma mais importante de consciência, a consciência *introspectiva*.

Na teoria dele, fenômenos mentais a respeito de algo (como, por exemplo, a percepção de objetos externos) são uma forma de consciência (justamente a consciência *perceptiva*), mas não a forma mais elaborada.[11] A forma mais desenvolvida da consciência (que viabilizaria uma interação mais complexa com o meio) é a consciência *introspectiva*, isto é, a percepção que um indivíduo tem dos seus próprios fenômenos mentais. Se a consciência de um fenômeno mental consiste em ele *ser percebido*, isto é, ser objeto de um fenômeno mental (perceptivo) de segunda ordem, isso faz da consciência de um fenômeno mental algo muito semelhante à percepção de um objeto ou de um estado de coisas no mundo extra psíquico, situação caracterizada pela *dualidade* entre a *observação* e o *objeto* observado.[12]

O objetivo do presente trabalho é oferecer uma *avaliação* da teoria de Armstrong a respeito da consciência e de sua concepção do inconsciente, de modo a elucidar tanto as suas *vantagens* quanto os seus próprios *defeitos*. Em primeiro lugar, será discutido em que sentido Armstrong defende uma perspectiva anti-

[10] A palavra "ciência", neste contexto, traduz a palavra "*awareness*", da língua inglesa, designando a característica de *estar ciente* de algo.

[11] A consciência perceptiva é uma forma de consciência *diferente* da consciência caracterizada na citação acima (cf. ARMSTRONG, 1968, p. 95), que é a consciência introspectiva. A consciência perceptiva é caracterizada por Armstrong como "consciência do que está ocorrendo no ambiente e no próprio corpo" (ARMSTRONG, 1997, p. 723). Assim sendo, como evidencia o exemplo do motorista de caminhão (exposto a seguir), é perfeitamente coerente que um episódio de consciência perceptiva seja *inconsciente* do ponto de vista da consciência introspectiva.

[12] Para Armstrong, quando há consciência dos próprios estados mentais, o que ocorre é que "*uma parte* do cérebro examina uma *outra parte* do cérebro. Na percepção, o cérebro examina o ambiente. Na ciência da percepção, outro processo no cérebro examina esse exame." (ARMSTRONG, 1968, p. 94 – grifos acrescentados). Em outra passagem, ele acentua que "Percepção é um assunto causal" (ARMSTRONG, 1997, p. 725), o que implica a *dualidade* entre causa e efeito que, neste caso, subjaz à dualidade entre o objeto (que afeta) e o sujeito (que é afetado por ele).

cartesiana, a saber: no sentido de que ele considera a consciência (de estados mentais) como um desenvolvimento *ulterior*, baseado em fenômenos mentais[13] pré-existentes (fenômenos que existem previamente e *independentemente* da consciência – cf. ARMSTRONG, 1997, p. 721) (seção 2). Em seguida, serão discutidas as noções, propostas por Armstrong, de (1) *consciência mínima* (a presença de algum tipo de atividade mental) e de (2) *consciência perceptiva* (a percepção – baseada em interações causais – de objetos e estados de coisas externos) (seção 3), bem como o conceito de (3) *consciência introspectiva* (a percepção dos próprios estados, eventos e processos mentais) (seção 4).

A tese defendida aqui é que, apesar do valor explicativo de seus três conceitos de consciência, e apesar de oferecer um quadro teórico que favorece nossa compreensão do inconsciente, Armstrong tem uma visão insatisfatória a respeito dos fenômenos mentais inconscientes, pois essa visão não dá conta do real *poder* do inconsciente em nossa vida psíquica (seção 5).

1 A PERSPECTIVA ANTI-CARTESIANA

Antes de mais nada, é importante perceber que Armstrong argumenta em favor de uma concepção anti-cartesiana da mente, no sentido de uma concepção na qual a consciência *não é* o aspecto essencial do mental, inclusive porque, segundo ele, parece perfeitamente razoável admitir que mesmo uma pessoa que estivesse *totalmente inconsciente* ainda teria uma mente, no sentido de ainda ser capaz de instanciar pelo menos algumas propriedades mentais. Armstrong afirma claramente que: "Existe, porém, uma tese sobre a consciência que eu creio poder ser rejeitada com confiança: a doutrina de Descartes de que a consciência é a essência da mentalidade." (ARMSTRONG, 1997, p. 721). Ele entende essa doutrina como baseada na assunção de que podemos explicar a mente em termos da consciência, mas ele acha que "a verdade vai na direção contrária. Na verdade, no sentido mais interessante da palavra 'consciência', a consciência é o creme sobre o bolo da mentalidade, um desenvolvimento especial e sofisticado da mentalidade. Ela não é o bolo em si" (ARMSTRONG, 1997, p. 721).

Evidentemente, tal perspectiva se encaixa perfeitamente com a aceitação de fenômenos mentais inconscientes, pois, como enfatiza Armstrong, ao recusar a consciência como o fundamento da mente, nós somos "forçados a admitir a possibilidade lógica de nos encontrarmos em um estado mental, mas não estarmos cientes de que nos encontramos nesse estado. Isso quer dizer, temos de admitir a possibilidade lógica de estados inconscientes." (ARMSTRONG, 1968, p. 113). Ele fundamenta essa possibilidade lógica de fenômenos mentais inconscientes alegando que "se a ciência introspectiva [*introspective awareness*] e seus objetos são 'existências distintas', como argumentamos, então tem que ser possível para os objetos que eles existam quando a ciência não existe" (Ibid, p. 114).

[13] Tais fenômenos mentais são entendidos por Armstrong como *idênticos* a estados neurofisiológicos do sistema nervoso central, cf. ARMSTRONG (1968, p. 89-90).

Mas Armstrong não se limita a admitir a simples possibilidade *lógica* da existência de fenômenos mentais inconscientes, "pois existem suficientes casos empíricos que podem ser interpretados naturalmente como implicando a existência atual de estados mentais dos quais não estamos cientes" (Ibid., p. 114), casos empíricos como os que serão discutidos a seguir, ao longo do presente trabalho (vide o caso do historiador, ou o caso do motorista de caminhão, discutidos abaixo). É em virtude de tudo isso que ele não tem reservas em considerar sua própria visão da mente como similar, em suas linhas básicas, àquela defendida por Freud:

> Em todos os momentos haverá estados e atividades de nossa mente dos quais nós não estamos introspectivamente cientes [*aware*]. Esses estados e atividades podem ser ditos estados e atividades mentais inconscientes em um bom sentido da palavra 'inconsciente' (ele é próximo do sentido freudiano, mas não há necessidade de sustentar que ele sempre envolve o mecanismo de repressão [*repression*]). (ARMSTRONG, 1997, p. 724).

Para compreender essa perspectiva a respeito da mente, devemos examinar as formas de existência mental independente de consciência, que são discutidas por Armstrong em relação aos três diferentes conceitos de consciência distinguidos por ele: (1) consciência *mínima*, (2) consciência *perceptiva* e (3) consciência *introspectiva*.

2 Consciência mínima e consciência perceptiva

Se a consciência não é algo constitutivo da mente, mas apenas "um desenvolvimento especial e sofisticado da mentalidade" (ARMSTRONG, 1997, p. 721), é importante compreender como ela se desenvolve a partir do alicerce formado pelos fenômenos mentais. De acordo com Armstrong, estados (eventos e processos)[14] mentais são aqueles que são *aptos* a provocar um comportamento de certo tipo (cf. ARMSTRONG, 1968, p. 82, p. 89) – sendo que ele entende que eles devem ser *identificados* com estados puramente físicos do sistema nervoso central (cf. Ibid., p. 89-90). No meu modo de entender, se consciência, de acordo com as três noções propostas por Armstrong, é algo de natureza *mental*, então tem que ser, na visão dele, algo *capaz* de provocar efeitos sobre o comportamento. Mas consciência, do modo como ele a entende, é algo que diz respeito a fenômenos mentais de natureza *perceptiva*: ele fala de "consciência" em termos de "reações comportamentais ao ambiente" (ARMSTRONG, 1997, p. 721), no sentido de que uma pessoa em um sono sem sonhos, ou sob anestesia total ainda possui alguma

[14] Armstrong enfatiza que seu uso do termo "estado" não é pensado de modo a excluir *eventos* e *processos* (cf. ARMSTRONG, 1968, p. 82). Ele entende que a noção de *estado* exprime a permanência de uma propriedade durante certo período, enquanto a noção de *processo* diz respeito a uma mudança que exige um certo tempo para ser completada (pois o estado existe inteiro durante o tempo em que uma propriedade permanece), e a noção de evento diz respeito ao surgimento ou desaparecimento de um estado (cf. Ibid., p. 130-131). Seguindo as caracterizações de Jaegwon Kim (1996, p. 06), podemos entender um *estado* como a instanciação continuada de uma propriedade, um *evento* como a mudança instantânea de uma propriedade para outra, e *processos* como sequências articuladas de estados e eventos.

ciência (*awareness*) mínima do ambiente,[15] "percepções do ambiente e do próprio corpo" (Ibid., p. 723), e em termos de "percepção do mental" (Ibid., p. 724). Se entendermos que "reações" ao ambiente supõem algum tipo de percepção, então o caráter perceptivo das três formas de consciência se torna claro, e já que a percepção é algum tipo de *atividade* mental, ou seja, algo que consiste em eventos e processos (e não em estados) mentais, então fica evidente o caráter *dinâmico* da consciência, nas três formas concebidas por Armstrong.

Se for assim, se torna compreensível porque Armstrong inicia suas considerações partindo do tipo de situação em que a consciência está (pelo menos aparentemente) ausente, como o caso de um sono sem sonhos: ele quer contrastar a situação em que existem eventos e processos mentais (como os processos perceptivos que constituem a consciência) com a situação em que existem apenas *estados* mentais (que constituem uma mentalidade *inerte*). Na ausência de qualquer atividade mental (como, p. ex., processos perceptivos) uma pessoa ainda possui mente no sentido de se encontrar em um imenso número de *estados* mentais. Armstrong dá o exemplo de um historiador especializado no período medieval: mesmo que ele estivesse em um sono sem sonhos, se nenhum evento ou processo mental ocorresse durante um certo período (ou seja, se não ocorresse nenhuma modificação psicológica), nós não negaríamos que ele possui um grande conjunto de conhecimentos e crenças acerca da idade média. Do mesmo modo, a uma pessoa totalmente inconsciente podem ser atribuídas memórias, habilidades, gostos, atitudes, emoções, desejos, propósitos etc. (que segundo uma concepção materialista da mente podem ser pensados como codificados fisicamente na estrutura do cérebro – cf. ARMSTRONG, 1997, p. 722; ARMSTRONG, 1968, p. 86-87). Por outro lado, existe uma série de atribuições mentais que nós não faríamos a uma pessoa totalmente inconsciente, como ter sensações, percepções ou explosões de desejo. Essa pessoa não pode pensar, contemplar ou se engajar em nenhum tipo de deliberação. O motivo pelo qual nós não faríamos essas atribuições é que sensações, percepções, explosões de desejo, pensamentos, contemplações e deliberações são *atividades* mentais, enquanto conhecimentos e crenças não são. A esse respeito o autor escreve:

> A distinção parece ser, aproximadamente em todo caso, a distinção entre eventos e ocorrências por um lado, e estados, por outro. Quando um estado mental produz efeitos mentais, o vir-a-ser [*comings-to-be*] de tais efeitos são eventos mentais: e assim atividade mental está envolvida. (ARMSTRONG, 1997, p. 722).

[15] Ao considerar que há uma ciência mínima, devido à presença de reações comportamentais, Armstrong parece estar considerando a "ciência" [*awareness*] em questão como algo semelhante àquilo que Gennaro chama de "ciência comportamental" [*behavioral awareness*], isto é: possuir estados internos, dotados de conteúdo proposicional, que são capazes de dirigir o comportamento da criatura em questão (cf. GENNARO, 1996, p. 06). Na visão dele, a ciência comportamental não implica ciência consciente, o que o leva a dizer, numa clara referência a Armstrong, que "o motorista de caminhão em longas distâncias está 'comportamentalmente ciente' das curvas na estrada" (Ibid., p. 06).

Armstrong propõe que, em situações onde alguma espécie de atividade mental[16] está ocorrendo seja usado o conceito de consciência *mínima* (*minimal consciousness*) – pensemos, por exemplo, no caso de pessoas que despertam sabendo a solução para um problema matemático, a qual eles desconheciam antes de ir dormir: parece necessário admitir que ocorreu atividade mental durante o sono (cf. ARMSTRONG, 1997, p. 722).

Neste ponto, Armstrong parece estar traçando uma diferença entre (a) a questão de um sujeito – ou um organismo – estar, ou não, desperto e (b) a questão de ele exemplificar, ou não, eventos e processos mentais. Para compreender melhor o conceito de consciência mínima proposto por Armstrong, é interessante considerar, mesmo que rapidamente, uma distinção proposta por outro importante filósofo da mente, David Rosenthal,[17] a saber, a distinção entre (*a*) *consciência de criatura* e (*b*) *consciência de estado*. De acordo com Rosenthal:

> Dois assuntos são frequentemente confundidos nas discussões sobre a consciência. Uma questão é: o que é para um *estado mental* ser consciente. Supondo que nem todos os estados mentais são conscientes, nós queremos saber como os estados conscientes se diferenciam daqueles que não são. E ainda que todos os estados mentais fossem conscientes, nós ainda perguntaríamos em que consiste a sua consciência. Denominamos essa a questão da *consciência de estado*. Esse será meu principal tema no texto que segue. Mas nós não descrevemos apenas estados mentais como sendo conscientes ou não; nós também atribuímos consciência a *criaturas*. Assim, existe uma segunda questão, a questão sobre o que é para uma pessoa ou outra criatura ser consciente, ou seja, como criaturas conscientes se diferenciam daquelas que não são conscientes. Denominamos esta a questão da *consciência de criatura.* (ROSENTHAL, 1997, p. 729; ROSENTHAL, 2017, p. 144 – grifos acrescentados).

Rosenthal entende que a questão da consciência de criatura, ou seja, a questão a respeito do que é para uma criatura, ou um organismo, estar consciente em dadas circunstâncias, não é um grande desafio para a reflexão filosófica, pois, segundo ele, nós dispomos de uma noção intuitiva do que é, em termos gerais, para um organismo estar consciente, a saber: tal organismo tem que estar *desperto* e *sensível* a seu ambiente circundante (cf. ROSENTHAL, 1997, p. 730; ROSENTHAL, 2017, p. 144).[18] Por isso ele elege a consciência de estado como o objeto de sua teoria da consciência.

E a consciência de estado, isto é, a consciência como uma *característica* de estados (eventos e processos) mentais, é pensada por Rosenthal em termos de

[16] Entendo que aquilo que Armstrong designa como "atividades mentais" corresponde claramente aos eventos e processos mentais na categorização de Kim (1996).

[17] Para uma discussão das três distinções referentes à consciência propostas por Rosenthal, cf. PRATA, 2017, p. 433-37.

[18] O influente filósofo australiano David Chalmers parece convergir com Rosenthal nesse ponto, quando ele afirma que a questão do que é para um organismo estar desperto é um tema para as ciências empíricas. Nas palavras de Chalmers: "Para uma explanação do sono e da vigília, será suficiente uma adequada explanação neurofisiológica dos processos responsáveis pelo comportamento contrastante do organismo nesses estados" (1995, p. 201).

nossa consciência *deles*. De acordo com ele: "estados conscientes são simplesmente estados mentais dos quais estamos conscientes de nos encontrar" (ROSENTHAL, 1986, p. 335), de modo que, como será detalhado posteriormente, a consciência como característica de um estado mental é pensada em termos da *relação* entre esse estado e o sujeito que o vivencia, no sentido de que esse sujeito "está transitivamente consciente daquele estado" (ROSENTHAL, 1997, p. 739; ROSENTHAL, 2017, p. 162).[19]

É interessante notar que, na teoria de Armstrong, haveria consciência mínima quando a consciência de criatura está ausente (ou seja, quando o organismo está desacordado) mas, mesmo assim, há algum tipo de atividade mental, no sentido de que, para além de simples *estados* (que não implicam nenhuma atividade), pelo menos algum *evento* ou *processo* mental tem lugar no psiquismo dessa criatura, o que, na visão materialista de Armstrong, significa apenas que certos processos *cerebrais* estão ocorrendo. E é importante notar também que, quando há consciência mínima, também a consciência de estado está ausente, pois as atividades mentais que constituem a consciência mínima são claramente, na descrição de Armstrong, atividades inconscientes. Portanto, a noção de consciência mínima proposta por ele se restringe a designar *atividades* mentais (idênticas a certos processos eletroquímicos no cérebro), que podem ter consequências no comportamento do organismo, mas permanecem *desprovidas* de consciência (no sentido de que o sujeito não tem consciência *delas*).

Nesse sentido, a noção de *consciência mínima* estabelece um contexto no qual a noção de *inconsciente* está envolvida, tanto no que diz respeito a criaturas, quanto no que diz respeito a seus estados (eventos e processos) mentais. Em outras palavras, a noção de consciência mínima implica, na verdade, *duas* noções de inconsciente (evidenciadas pela primeira distinção de Rosenthal).

Portanto, fica claro que essas situações, nas quais ocorre aquilo que Armstrong chama de "consciência mínima", certamente não são nosso referencial quando nos referimos ao fenômeno da consciência, pois a consciência mínima ocorre em circunstâncias que, cotidianamente, são chamadas de "inconsciência". Quando falamos de consciência, levamos em consideração outros fatores, e ele destaca que entre esses fatores, está a capacidade de *percepção* por parte do ser em questão. Consciente, em um sentido importante da palavra, é o ser capaz de perceber o que se passa ao seu redor e no seu próprio corpo.[20] Se um indivíduo humano, p. ex., está dormindo, mas alguma atividade mental está em curso, então ele tem o que Armstrong chama de consciência mínima, mas ainda lhe falta algo para que se atribua a ele consciência no sentido usual da palavra. A partir do momento em que o indivíduo seja novamente capaz de percepção, algo importante

[19] O conceito de consciência *transitiva* será explicitado a seguir.

[20] "Entre as atividades mentais, porém, parece que nós fazemos uma ligação especial entre consciência e *percepção*. Na percepção, existe consciência do que está ocorrendo atualmente no ambiente de alguém e em seu próprio corpo (claro que a consciência pode envolver ilusão). Existe um sentido importante no qual se uma pessoa não está percebendo, então não está consciente, mas se percebe, então está consciente" (ARMSTRONG, 1997, p. 723). Esta passagem deixa claro o caráter *intencional* da consciência perceptiva, ou seja, o seu caráter de ser *acerca* ou *a respeito de* algo.

foi acrescentado, aquilo que o autor chama de consciência *perceptiva* (*perceptual consciousness*).

É interessante notar que a distinção entre (*a*) consciência de criatura e (*b*) consciência de estado, proposta por Rosenthal, também é relevante para a compreensão da consciência perceptiva, tal como concebida por Armstrong, pois, por um lado, a consciência perceptiva costuma acontecer quando a consciência de criatura está presente (embora o exemplo do motorista de caminhão, discutido abaixo, poderia levar alguém a argumentar que a consciência perceptiva pode acontecer quando a consciência de criatura está ausente) e, por outro lado, a consciência perceptiva ocorre quando estão presentes certos estados mentais, justamente as *percepções*, do próprio corpo ou do ambiente circundante.[21] Nesse sentido, essa distinção de Rosenthal ajuda a delinear certos aspectos presentes na noção de consciência perceptiva de Armstrong.

Mas além disso, há uma segunda distinção proposta por Rosenthal, a distinção entre (a') *consciência intransitiva* e (b') *consciência transitiva*, e ela também é pertinente para elucidar o conceito de consciência perceptiva, pois a percepção é um exemplar de consciência transitiva. De acordo com Rosenthal:

> Colocando a consciência de criatura de lado, podemos distinguir duas maneiras como usamos a palavra 'consciente'. Uma é quando falamos de nosso estar consciente *de* alguma coisa. Por causa do objeto direto, devo chamar esse uso de *transitivo*. Mas nós também aplicamos o termo 'consciente' a estados mentais, para dizer que eles são estados conscientes. Isso é o que eu rotulei como consciência de estado. A falta de um objeto direto sugere chamar esse uso de *intransitivo*. Esse uso intransitivo tem lugar somente quando falamos de estados mentais, ao passo que nós falamos de estar consciente, transitivamente, tanto *de* coisas físicas quanto mentais. Nós podemos estar transitivamente conscientes *de* uma pedra, *de* uma sinfonia, ou *de* um estado mental. (ROSENTHAL, 1997, p. 737; ROSENTHAL, 2017, p. 158-159 – grifos acrescentados).

Nesta passagem, Rosenthal está tratando de algo que já foi mencionado anteriormente, na introdução do presente artigo (cf. também GENNARO, 1996, p. 03), a saber: ele está tratando da distinção entre (a') consciência enquanto uma propriedade expressa por um predicado *monádico* (p. ex. "... é consciente", que é um predicado intransitivo) onde o lugar vazio pode ser preenchido com a designação de uma criatura ou de um estado mental e (b') consciência enquanto uma propriedade expressa por um predicado *diádico* (p. ex. "... é consciente de ..." que é um predicado transitivo) onde o primeiro lugar vazio tem que ser substituído pela designação de uma criatura[22] e o segundo lugar vazio pode ser preenchido com a designação de um fenômeno físico ou de um fenômenos

[21] Sendo uma percepção um estado mental, coloca-se a pergunta sobre se esse estado é, em dado momento, consciente, ou não.

[22] Rosenthal afirma claramente que "Ser transitivamente consciente de algo é uma relação na qual *uma pessoa ou outra criatura* está com esse algo. Assim, apenas criaturas podem estar transitivamente conscientes de coisas" (ROSENTHAL, 1997, p. 738; ROSENTHAL, 2017, p. 160 – grifo acrescentado). A esse respeito, cf. Kriegel (2009, p. 27).

mental.[23] O ponto a ser destacado é que a consciência perceptiva de Armstrong é um tipo específico daquilo que Rosenthal denomina "consciência transitiva". Como esclarece Rosenthal:

> Podemos chamar esse fenômeno de *consciência transitiva*. Alguém está consciente *de* algo quando vê ou escuta esse algo, ou o sente e percebe de algum outro modo. Ter um pensamento sobre algo algumas vezes também é suficiente para se ter consciência desse algo, mas não sempre. (ROSENTHAL, 2008a, p. 239).[24]

Entretanto, a distinção entre (a') consciência intransitiva e (b') consciência transitiva envolve uma sutileza à qual precisamos prestar muita atenção, pois essa sutileza é decisiva para que possamos compreender o próximo conceito de consciência proposto por Armstrong. O ponto é: no que uma percepção é percepção *de* algo ela é uma forma de consciência transitiva, a saber, uma consciência perceptiva *de* um objeto. Nesse sentido, a percepção desfruta da característica que o debate filosófico contemporâneo, desde Brentano, chama de *Intencionalidade*,[25] característica esta que mesmo as percepções que não são objeto de introspecção (ou seja, as percepções inconscientes) possuem.[26] Por outro lado, se a percepção de um objeto externo é consciente – e a tradição cartesiana argumenta que ela é *necessariamente* consciente – então, se trata de *consciência intransitiva*, pois apesar do fato de a percepção ser transitiva, ela pode ser descrita de uma perspectiva na qual podemos atribuir *a ela* (a percepção) uma propriedade expressa por um predicado monádico.

Se, por exemplo, digo que "minha percepção do pôr do sol é consciente", (no sentido de que eu estou consciente *da* percepção – cf. ROSENTHAL, 1986, p. 335; ROSENTHAL, 1997, p. 739; ROSENTHAL, 2017, p. 161) a consciência, nessa frase entre aspas, está sendo apresentada como uma propriedade não relacional dessa percepção, e essa propriedade não relacional é o que Rosenthal chama de consciência *intransitiva*. Para perceber isso com mais clareza, basta trocar a descrição "percepção *do* pôr do sol" (descrição que expressa uma consciência transitiva) pela abreviação "P", pois se digo que "P é consciente", embora P envolva

[23] Como coloca Rosenthal: "Nós podemos estar transitivamente conscientes de uma pedra, de uma sinfonia, ou de um estado mental" (ROSENTHAL, 1997, p. 737; ROSENTHAL, 2017, p. 159).
[24] Não sempre porque apenas pensamentos *assertóricos* são capazes de tornar consciente o estado mental a respeito do qual eles são, e nem todo pensamento é assertórico.
[25] Na mais famosa passagem de sua *Psicologia do ponto de vista empírico*, publicada originalmente em 1874, Brentano escreveu: "Todo fenômeno psíquico é caracterizado por aquilo que os escolásticos da idade média denominaram o 'existir em' [*Inexistenz*] intencional [...], o que nós denominaríamos, embora não com expressões totalmente precisas, a relação a um conteúdo, a direção a um objeto (com o que não se deve entender aqui uma realidade), ou a objetividade imanente. Todo fenômeno psíquico contém em si algo como objeto, embora não todos do mesmo modo. Na representação algo é representado, no juízo algo é reconhecido ou recusado, no amor amado, no ódio odiado, no desejo desejado, etc." (BRENTANO, 1924, p. 124-125; BRENTANO, 1995, p. 88). Se percebo o por do sol, essa minha percepção efetiva uma consciência *de*, ou seja, uma consciência *direcionada* a um certo estado de coisas.
[26] É interessante notar que Armstrong reconhece a tese de Brentano de que a Intencionalidade é a marca do mental, por mais que ele não aceite que isso signifique uma diferença entre o mental e o físico. De acordo com ele: "É claro que nenhum fisicalista pode aceitar a irredutibilidade da intencionalidade, embora ele possa aceitar a visão de Brentano de que a intencionalidade é a marca do mental" (ARMSTRONG, 1968, p. 41).

uma *relação* com o seu objeto (em nosso exemplo, uma relação intencional com o sol), o que está sendo expresso é uma propriedade intransitiva, expressa por um predicado *monádico*.[27]

Mas por que essa sutileza é importante? Porque, como vimos, Armstrong *rejeita* a visão cartesiana e defende que estados mentais, mesmo os que são consciência transitiva de um objeto, podem existir independentemente da consciência. E isso significa que esses estados mentais, no que são *inconscientes*, existem desprovidos de consciência *intransitiva*, ou seja, eles existem desprovidos da propriedade expressa pelo predicado monádico "... é consciente".

A esse respeito, é importante perceber que, apesar de a expressão "consciência inconsciente" parecer, à primeira vista, contraditória, ela pode ser empregada de uma maneira que é perfeitamente razoável (se ela for construída com dois sentidos distintos de "consciência"). Uma maneira de dissipar a impressão de contradição é substituir o termo "consciência" (*consciousness*) pelo termo "ciência" (*awareness*) – formando, então a expressão "ciência inconsciente" – pois, como esclarece Gennaro:

> É claro que 'ciente' (*aware*) e 'consciente' (*conscious*) não são meramente sinônimos [...]. Não é contraditório falar de ser 'inconscientemente ciente' de algo. De modo similar, a expressão 'conscientemente ciente' não é redundante. Ciência não necessariamente carrega conotações de consciência. (GENNARO, 1996, p. 05).

Para perceber isso, basta pensar que eu, por exemplo, estou sempre ciente de que nasci em uma cidade localizada ao sul do equador, embora muito raramente eu pense sobre isso de forma consciente. Sendo assim, é perfeitamente razoável dizer que estou, durante a maior parte do tempo, *inconscientemente ciente* desse fato.

Mas voltando para a questão da consciência como propriedade de estados mentais, é importante perceber que o que garante que estados mentais, eventualmente, desfrutem daquilo que Rosenthal chama de "consciência intransitiva", é uma terceira forma de consciência, que não se confunde com (1) a consciência mínima nem com (2) a consciência perceptiva de objetos externos (embora, na verdade, seja uma forma específica de consciência perceptiva, como veremos). A questão é que, para Armstrong, mesmo se há (1) *atividade mental ocorrendo*, e se essa atividade mental inclui também (2) *percepção genuína* (e não

[27] Mas é importante ter em vista que essa propriedade intransitiva, como já foi aludido acima, é pensada por Rosenthal como *decorrente* de uma relação: a relação com o sujeito. De acordo com ele: "a consciência intransitiva de um estado mental é simplesmente uma certa maneira do indivíduo estar transitivamente consciente daquele estado" (ROSENTHAL, 1997, p. 739; ROSENTHAL, 2017, p. 162). Embora não seja um defensor da perspectiva de ordem superior, o filósofo Uriah Kriegel argumenta a favor da plausibilidade da tese de que o *caráter subjetivo* de um estado mental consciente (caráter em virtude do qual esse estado se manifesta ao sujeito) implica uma consciência *do* estado mental, pois "seria, de fato, bastante estranho manter que uma experiência é *para* o sujeito ainda que o sujeito não tenha nenhuma ciência dela. Uma vez que essa ciência [*awareness*] é ciência-de, ela envolve uma relação transitiva [*of-ness relation*] com a experiência" (KRIEGEL, 2009, p. 104).

o tipo de experiência que ocorre em um sonho ou em uma alucinação), ainda há algo importante que pode estar ausente: aquilo que Armstrong chama de (3) *consciência introspectiva*.

3 Consciência introspectiva

Se a consciência perceptiva foi caracterizada acima como uma consciência *dos acontecimentos e objetos do ambiente* (cf. ARMSTRONG, 1968, p. 95; ARMSTRONG, 1997, p. 723), a consciência introspectiva, evidentemente, pode ser caracterizada como uma capacidade através da qual "nos tornamos cientes das ocorrências que estão tendo lugar em nossa própria mente" (ARMSTRONG, 1968, p. 95). Ao caracterizar a consciência introspectiva como uma "percepção do mental" ou um "sentido interno" (ARMSTRONG, 1997, p. 724), Armstrong está ciente de estar levando adiante uma antiga tradição filosófica, que ele identifica em autores como Locke e Kant (cf. ARMSTRONG, 1968, p. 95; ARMSTRONG, 1997, p. 724), tradição que, aliás, é muito mais antiga do que esses autores, remontando aos primórdios da filosofia ocidental. Para explicar do que se trata, o autor apresenta uma situação – bastante conhecida na literatura em filosofia da mente – e já vivenciada por diversas pessoas[28] em situações peculiares:

> Depois de dirigir por longos períodos de tempo, particularmente à noite, é possível 'chegar' [*to 'come to'*] e se dar conta de que por algum tempo a pessoa estava dirigindo sem estar ciente [*aware*] do que estava fazendo. Se dar conta disso é uma experiência alarmante. É natural descrever o que aconteceu antes da pessoa se dar conta dizendo que durante aquele período ela carecia de consciência. (ARMSTRONG, 1997, p. 723).

Nesse exemplo, parece bastante claro que, nos dois sentidos propostos por Armstrong que discutimos até agora, consciência tinha lugar, pois havia (1) *atividade mental* e também (2) *percepção*, como prova a sequência de tarefas bastante complexas que o indivíduo executou naquele período, afinal ele dirigiu o veículo ao longo da estrada (talvez por vários quilômetros durante alguns minutos). Se ele tivesse parado de perceber o ambiente ao seu redor certamente um acidente teria ocorrido. Armstrong afirma que é necessário admitir, no mínimo, que nessa situação os olhos e o cérebro têm que ter sido estimulados exatamente da mesma maneira que em casos comuns de percepção. "Por que então negar que percepção ocorreu?" (ARMSTRONG, 1997, p. 723).[29] Mas mesmo se admitirmos que havia consciência mínima e consciência perceptiva, parece claro que algo mais

[28] Embora eu mesmo nunca tenha vivenciado algo assim, conheço diversas pessoas que relatam experiências semelhantes a essa mencionada por Armstrong.

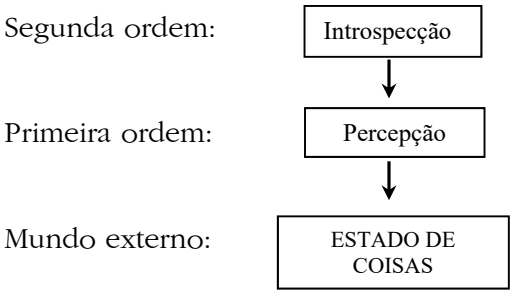
[29] É possível que muitos (provavelmente simpatizantes da visão cartesiana) negassem que se trata de percepção propriamente dita, na medida em que nenhuma *vivência* (*experience*) teve lugar, isto é, na medida em que não estava presente nenhum *aspecto fenomenológico*. Mas o ponto de Armstrong parecer ser o de que algo ocorreu no sujeito que desempenhou a *função* que a percepção desempenharia em uma situação normal, de modo que isso que ocorreu nesse caso atípico mereceria ser chamado de "percepção".

estava ausente, aquilo que o autor considera o sentido mais interessante da palavra "consciência".

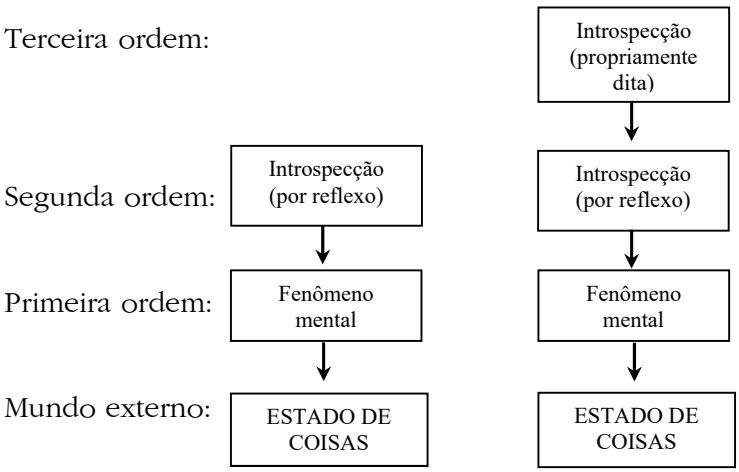
O caso do motorista de caminhão pode parecer extraordinário (dirigir o veículo por algum tempo sem ter tido ideia do que estava fazendo), mas, para Armstrong, tal caso diz respeito a um nível relativamente simples, em termos evolucionários, de atividade mental: ação voltada para um objetivo e guiada por percepção. Ele especula que talvez muitos animais, que têm um sistema nervoso menos desenvolvido do que o nosso, vivam continuamente (ou pelo menos a maior parte do tempo) no estado em que o motorista ficou durante alguns minutos. Talvez a terceira forma de consciência que estava ausente seja um desenvolvimento evolucionário tardio. Armstrong caracteriza essa forma de consciência como se segue:

> O que faltava ao motorista de caminhão de longa distância? Eu penso que é uma forma adicional de percepção ou, um pouco mais cautelosamente, é algo que se assemelha à percepção. Mas diferente da percepção sensorial, ela não é dirigida ao nosso ambiente e/ou ao nosso estado corporal em curso. É percepção do mental. Tal percepção 'interna' é tradicionalmente chamada de introspecção, ou ciência [*awareness*] introspectiva. Nós podemos, portanto, chamar esse terceiro tipo de consciência de consciência 'introspectiva'. (ARMSTRONG, 1997, p. 724).

A consciência introspectiva, claramente, implica consciência mínima (a presença de algum tipo de atividade mental). Mas, de acordo com Armstrong, se o conceito de consciência perceptiva for restringido à percepção do exterior e do próprio corpo através dos sentidos, então a consciência introspectiva não implica consciência perceptiva (o que, no meu entendimento, significa que a consciência introspectiva não tem necessariamente que ser a respeito de percepções – pode ser também a respeito de memórias, emoções, desejos, etc.). A consciência introspectiva é, na concepção dele, uma ciência (à maneira da percepção) dos estados e atividades em curso na própria mente de um indivíduo, sendo que tais atividades *podem* incluir a percepção sensorial (do mundo externo e do próprio corpo), mas podem também não incluir. Podemos representar a consciência introspectiva de uma percepção (que é um tipo entre diferentes outros tipos de fenômenos mentais) através da seguinte figura:



Interessante é que a consciência introspectiva, na medida em que é uma atividade mental, pode se tornar ela própria objeto de uma introspecção ulterior. É importante notar que Armstrong distingue entre uma introspecção *espontânea*, responsável pela consciência de todo e qualquer estado mental efetivamente consciente, e uma introspecção *deliberada*, que é justamente, essa introspecção da introspecção. Nesse sentido, é imprescindível perceber que a consciência introspectiva, tal como foi discutida acima, *não é*, em primeira linha, o esforço consciente de examinar o que se passa na própria mente, mas sim uma percepção interna, em certo sentido, espontânea, não planejada, que efetiva a consciência de um estado mental. Armstrong (1997, p. 725) denomina essa introspecção de "consciência introspectiva por reflexo" [*"reflex" introspective consciousness*]. Já o esforço consciente de escrutinar os próprios fenômenos mentais é denominado por ele de "introspecção propriamente dita" [*"introspection proper"*], algo que consistiria em uma percepção de terceira ordem que se volta para a percepção de segunda ordem que torna um dado fenômeno mental consciente. Trata-se de uma percepção de terceira ordem porque, assim como na introspecção por reflexo há "um outro estado mental, dirigido aos estados internos originais" (ARMSTRONG, 1968, p. 94), na introspecção propriamente dita tem que haver um outro estado mental que esteja dirigido ao segundo estado.[30] A diferença entre essas duas formas de introspecção pode ser representada através da seguinte figura:



A esse respeito, penso que a terminologia de Armstrong é um pouco confusa, pois o termo "introspecção" sugere um esforço deliberado de autoexame (como o próprio Armstrong parece sugerir ao falar de uma "introspecção propriamente dita"), de maneira que o uso desse termo para designar a consciência, por assim dizer, espontânea de certos estados mentais, pode induzir a erro. Naquilo que Armstrong chama de "consciência introspectiva por reflexo", o que temos é o *monitoramento* de um estado mental por outro estado mental do mesmo indivíduo, e ele escolhe designar essa relação entre dois estados mentais

[30] Para uma discussão do conceito de introspecção, e dos diferentes níveis de estados mentais envolvidos, cf. Rosenthal (1986, p. 337).

com a palavra "introspecção". Mas eu entendo que seria mais adequado respeitar o sentido específico dessa palavra,[31] e reservá-la para aquilo que Armstrong chama de "introspecção propriamente dita", pois quando um estado mental está consciente (isto é, exemplifica a característica *intransitiva* de ser consciente) de modo espontâneo ("por reflexo", nos termos de Armstrong), a atenção do indivíduo está voltada para aquilo que é o tema desse estado mental. Neste ponto, a teoria de Armstrong pode ser iluminada através de uma terceira distinção conceitual operada por Rosenthal, a saber, a distinção entre consciência não introspectiva, por um lado, e consciência introspectiva, por outro. De acordo com Rosenthal:

> Quando prestamos atenção *deliberadamente* ao estado mental no qual nos encontramos, estamos introspectivamente conscientes desse estado. Isso é diferente do modo como estados mentais são conscientes quando não estamos deliberadamente focando nossa atenção sobre eles. Introspecção é consciência atenta e deliberada de nossos estados mentais. É algo relativamente raro, e é algo mais elaborado do que a maneira na qual estados mentais são ordinariamente conscientes. (ROSENTHAL, 1997, p. 730; ROSENTHAL, 2017, p. 145 – grifo acrescentado).

As peculiaridades da introspecção se obscurecem quando se chama "introspectiva" essa consciência mais simples dos próprios estados mentais. Mas provavelmente Armstrong escolheu esse termo (que me parece pouco adequado) porque ele sugere um exame, ou um *monitoramento* da própria mente do sujeito que se encontra em estados mentais conscientes, ou seja, uma *relação* do sujeito com os seus próprios estados mentais.[32]

O caso é que, assim como a percepção sensorial não é uma atenção completa a todos os estados e processos em curso no ambiente e no próprio corpo, aquilo que Armstrong chama de consciência introspectiva não abrange a totalidade dos estados e processos em curso na mente de um indivíduo. Nas palavras dele:

> Em todos os momentos haverá estados e atividades de nossa mente dos quais nós não estamos introspectivamente cientes [*aware*]. Esses estados e atividades podem ser ditos estados e atividades mentais inconscientes em um bom sentido da palavra 'inconsciente' (ele é próximo do sentido freudiano, mas não há necessidade de sustentar que ele sempre envolve o mecanismo de repressão [*repression*]). Tais estados e atividades mentais inconscientes podem naturalmente envolver consciência mínima e/ou consciência perceptiva, na verdade as *atividades* envolvem consciência mínima por definição. (ARMSTRONG, 1997, p. 724).

[31] A diferença entre a consciência introspectiva e a simples consciência de estados mentais é elucidada por Gilbert Ryle nos seguintes termos: "Introspecção é uma operação atenta e que é realizada apenas ocasionalmente, enquanto que a consciência é suposta como um elemento constante de todos os processos mentais, e cujas revelações não exigem que sejam recebidas em atos especiais de atenção. Além disso, nós realizamos introspecção com a intenção de encontrar respostas para problemas específicos, enquanto que estamos conscientes quer desejemos ou não; todo mundo está constantemente consciente, enquanto desperto, mas somente aqueles que estão, de tempos em tempos, interessados no que ocorre em suas mentes realizam introspecção" (RYLE, 1970, p. 157).

[32] Para conhecer os argumentos de Armstrong a favor de seu próprio uso do termo "introspecção", cf. Armstrong (1968, p. 95).

E exatamente como a percepção externa não abrange tudo o que ocorre no ambiente, ela também é, além disso, *potencialmente falha*, pois pode não corresponder com exatidão à realidade. Do mesmo modo, também a consciência introspectiva pode falhar em corresponder com exatidão à realidade mental, o que leva Armstrong a rejeitar a tese cartesiana da indubitabilidade da consciência.[33] É interessante notar que, ao considerar a introspecção potencialmente falha, como é a percepção externa, o autor, sugere a presença da dicotomia entre conhecimento e objeto *dentro* da própria mente. Da mesma maneira, quando considera uma concepção *causal* da introspecção a concepção mais plausível para um filósofo naturalista (cf. ARMSTRONG, 1997, p. 725), isto é, quando supõe que fenômenos mentais exercem poder casual sobre a introspecção (assim como fazem objetos externos sobre a percepção), Armstrong trata os fenômenos mentais como algo essencialmente separado da consciência introspectiva que se pode ter deles, o que parece fazer da introspecção uma relação *cognitiva*, uma forma de *conhecimento*.

Armstrong entende que a função biológica da consciência introspectiva é tornar nossa atividade mental mais *sofisticada*, de modo que essa atividade possa propiciar ações mais complexas. Uma vez que dispomos de uma capacidade perceptiva interna, que nos torna cientes [*aware*] dos estados e atividades mentais que estão em curso em nossa mente, se torna muito mais fácil *integrar* esses estados e atividades, ou seja: "colocá-los para trabalhar juntos de maneiras complexas e sofisticadas que são necessárias para se alcançar fins complexos e sofisticados" (ARMSTRONG, 1997, p. 726). Se a função biológica dos processos mentais é mediar os estímulos ambientais e a resposta do organismo, de modo a tornar esta resposta mais sofisticada e eficiente, a função biológica da introspecção é informar o agente daquilo que se passa em seu psiquismo, permitindo que ele trabalhe de modo mais eficaz os estímulos externos (e, eventualmente *internos* ao próprio organismo) e suas próprias reações a eles (cf. ARMSTRONG, 1968, p. 163).

Como exemplo do primeiro caso, podemos pensar em alguém confrontado com uma situação problema (como um cirurgião diante de uma súbita complicação em uma cirurgia, ou um investidor em dúvida sobre vender um ativo). Nessas circunstâncias, para Armstrong, "várias respostas possíveis podem ser tentadas 'na imaginação' para ver qual resposta vai satisfazer melhor os propósitos do agente. Como um resultado causal, a resposta pode ser muito mais eficiente" (ARMSTRONG, 1968, p. 163). Como exemplo do segundo caso (relativo a estímulos internos), Armstrong cita a situação em que alguém tem que lidar com a própria agressividade. De acordo com ele: "conhecimento da presença, dentro de nós, de potenciais causas de comportamento, é de um valor óbvio na condução da vida" (Ibid., p. 99).

[33] A referência a Descartes sugere que Armstrong tem em vista nesta passagem aquilo que ele chama "introspecção propriamente dita" (1997, p. 725), que me parece corresponder ao que Rosenthal chama de "consciência introspectiva". Penso que essas considerações de Armstrong poderiam ser entendidas como uma refutação da *incorrigibilidade* afirmada por Descartes, mas não da *indubitabilidade*, pois, para o filósofo francês, o que é indubitável é o fato do "eu penso" e não os conteúdos do pensamento (que parecem ser o que Armstrong tem em mente com essas considerações).

Considerações finais

A discussão empreendida no presente trabalho, no meu modo de entender, mostra que a teoria de Armstrong sobre a consciência se revela interessante e fecunda para se pensar a respeito desse complexo conjunto de fenômenos, uma vez que as distinções propostas por ele (entre consciência (a) *mínima*, (b) *perceptiva* e (c) *introspectiva*) parecem tornar discerníveis aspectos importantes da consciência, e a partir do momento em que somos capazes de enxergar essas diferenças, nosso entendimento da consciência se amplia. Por mais que (i) desconsidere uma distinção importante acerca da consciência, como é a distinção entre consciência de *criatura* e consciência de *estado*, e por mais que (ii) não trace, com suficiente clareza, uma distinção bastante elucidativa, que é aquela entre consciência *não introspectiva* e consciência *introspectiva* (o que revela duas limitações importantes); a conceituação proposta por Armstrong se mostra bastante proveitosa para nossa compreensão da temática da consciência, uma vez que ela nos ajuda a discernir entre aspectos muito relevantes, como o são (a) as *atividades mentais* que, de algum modo, nos conectam com o mundo, (b) as *conexões causais* que nos tornam perceptivamente cientes do nosso ambiente imediato, e (c) nossa conexão com nossos *próprios fenômenos mentais*, que efetivam nossas relações psicológicas com o mundo (e com nós mesmos). Entendo que, apesar de suas limitações, os conceitos de consciência elaborados por Armstrong têm considerável valor para a reflexão teórica acerca dos diversos fenômenos que designamos com a palavra "consciência".

Além disso, uma teoria como essa representa um grande avanço em relação à tradição cartesiana, que entende a consciência como *essencial* aos fenômenos mentais, tradição que continuou influente no século XX através de autores como Brentano[34] e Sartre,[35] e que continua a reverberar na atual filosofia da mente.[36] A visão cartesiana da consciência se mostra muito inadequada para explicar os abundantes indícios de que há toda uma vasta vida mental inconsciente (cf. ROSENTHAL, 1986, p. 329, p. 334; ROSENTHAL, 1997, p. 731; ROSENTHAL, 2017, p. 147-148; BARGH & MORSELLA, 2008), pois essa visão está comprometida com uma teoria *disposicional* do inconsciente (cf. DESCARTES, *Respostas às quartas objeções* [AT, VII, p. 246; AT, IX, p. 190];[37] SEARLE, 1992, p. 161; SEARLE, 1997, p. 231), teoria que padece de diversos problemas sérios (cf. ROSENTHAL, 1986, p.

[34] Em sua obra *Psicologia do ponto de vista empírico*, Brentano procurou refutar os argumentos que ele identificou na literatura pertinente a favor da existência de fenômenos mentais inconscientes. Cf. Brentano (1924, p. 147-148); Brentano (1995, p. 105).

[35] Sartre considerava que "a existência de um fenômeno psíquico e o sentido que ele tem para a consciência são uma coisa só" (SARTRE, 1996, p. 36), posicionamento que, evidentemente, o obrigava a "rejeitar inteiramente a existência de um inconsciente" (Ibid., p. 36, nota nº 13). Sobre as perspectivas de Brentano e Sartre a respeito da consciência, cf. PRATA (2016).

[36] Embora alegue divergir da perspectiva cartesiana (cf. SEARLE, 1992, p. 13-14; SEARLE, 1997, p. 24-25; SEARLE, 2004, p. 13), John Searle defende abertamente a tese de que "todas as outras noções mentais – como intencionalidade, subjetividade, causação mental, inteligência, etc. – só podem ser plenamente compreendidas como *mentais* por meio de suas relações com a consciência" (SEARLE, 1992, p. 84; SEARLE, 1997, p. 125-126).

[37] Esse texto não está disponível na edição brasileira consultada para a redação do presente trabalho.

342; ROSENTHAL, 1997, p. 732; ROSENTHAL, 2017, p. 149; PRATA, 2018, p. 515-20).

Nesse sentido, uma teoria segundo a qual os fenômenos mentais existem *independentemente* de nossa consciência deles (cf. ARMSTRONG, 1968, p. 114) oferece um quadro teórico claro, dentro do qual podemos compreender o estatuto dos fenômenos mentais inconscientes em nossa vida mental.

Todavia, as concepções perceptivas da consciência enfrentam as suas próprias dificuldades (cf. ROSENTHAL, 1997, p. 739-740; ROSENTHAL, 2017, p. 163-164; VAN GULICK, 2012, p. 48). E, considerando a perspectiva adotada no presente trabalho, torna-se importante ressaltar que a maneira como Armstrong entende o *inconsciente* padece de algumas deficiências, pois uma vez que ele atribui à consciência um *incremento* das capacidades mentais, os fenômenos inconscientes aparecem em sua teoria de uma maneira distorcida, como se eles fossem muito menos poderosos do que eles realmente são.

O ponto é que ao pensar a consciência como responsável por uma *integração* de diversos processos mentais, integração que os tornaria mais *sofisticados* e *eficientes*, Armstrong sugere que os fenômenos mentais inconscientes possuem apenas capacidades bastante limitadas. Por exemplo, ele defende que a capacidade de solução de problemas depende da capacidade de *pensar conscientemente* sobre as diversas respostas possíveis, e da capacidade de *escolher conscientemente* a resposta a um problema que melhor se adéqua aos próprios objetivos, de modo que o animal capaz de resolver problemas mentalmente precisa ter ciência dos estados mentais relevantes (cf. ARMSTRONG, 1968, p. 163; ROSENTHAL, 2008b, p. 831).

O problema é que, se uma visão como essa estava em consonância com a perspectiva que, até algum tempo atrás, era predominante na psicologia cognitiva (cf., GREENWALD, 1992; LOFTUS & KLINGER, 1992), tal visão, simplesmente, não resiste aos resultados de estudos mais recentes. Bargh & Morsella (2008, p. 73) enfatizam que, diferente da psicologia cognitiva, as pesquisas em psicologia social, ao longo dos últimos 30 anos, consolidaram o ponto de vista de que a mente inconsciente é uma influência constante e poderosa sobre os processos mentais superiores.

O psicólogo holandês Ap Dijksterhuis (2004, p. 587) chama a atenção para o fato de que a atividade mental consciente tem uma capacidade relativamente baixa de processamento de informação (algo em torno de 40 a 60 bits por segundo, o equivalente a uma pequena sentença linguística), e destaca que os resultados de diversos estudos empíricos, realizados desde a década de 1990, mostram que a atividade consciente tem capacidades reduzidas para a solução de problemas. E ele próprio realizou um conjunto de experimentos sobre a tomada de decisões que reforçam essa conclusão. De acordo com ele:

> Pensamento inconsciente melhorou a qualidade das decisões. Quando os sujeitos foram confrontados com decisões complexas, alguns minutos de distração, durante os quais eles puderam se engajar em pensamento inconsciente – mas não em pensamento consciente – levaram a decisões superiores se comparadas a

> circunstâncias nas quais os sujeitos não puderam se engajar em pensamento inconsciente, ou a circunstâncias nas quais eles se engajaram em pensamento consciente. (DIJKSTERHUIS, 2004, p. 596).

Portanto, a despeito de seus méritos para a elucidação do fenômeno da consciência, e para a articulação de uma concepção do inconsciente, a teoria de David Armstrong não estava à altura do verdadeiro papel dos fenômenos mentais inconscientes em nossa vida psicológica.

Referências

ADAM, Charles; TANNERY, Paul. (Eds.). *Ouvres de Descartes – Meditationes de prima philosophia*. Paris: Vrin, 1996. (Vol. VII).

_____. *Ouvres de Descartes – Méditations et príncipes* (traduction française). Paris: Vrin, 1996. (Vol. IX).

ARMSTRONG. David. *A materialist theory of mind*. London: Routledge & Kegan Paul, 1968.

_____. What is Consciousness? In: BLOCK, Ned; FLANAGAN, Owen; GÜZELDERE, Guven (Eds.). *The nature of consciousness*: philosophical debates. Cambridge, MA: MIT Press, 1997. p. 721-728.

BARGH, John A.; MORSELLA, Ezequiel. The unconscious mind. *Perspectives on Psychological Science*, v. 3, n. 1, p. 73-79, 2008.

BLOCK, Ned. On a confusion about a function of consciousness. In: BLOCK, Ned; FLANAGAN, Owen; GÜZELDERE, Guven (Eds.). *The nature of consciousness*: philosophical debates. Cambridge, MA: MIT Press, 1997. p. 375-415.

BRENTANO, Franz. *Psychologie vom empirischen Standpunkt*. 2.ed. Leipzig: Felix Meiner, 1924.

_____. *Psychology from an empirical standpoint*. London: Routledge & Kegan Paul, 1995.

CAROPRESSO, Fátima. Inconsciente, cérebro e consciência: reflexão sobre os fundamentos da metapsicologia freudiana. *Scientiae Studia*, v. 7, n. 2, p. 271-82, 2009.

CATALDO-MARIA, Thiago. M. S.; WINOGRAD, Monah. Freud e Brentano: mais que um flerte filosófico. *Psico*, v. 44, n. 1, p. 34-44, 2013.

CHALMERS, David. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, v. 2, n. 3, p. 200-219, 1995.

DESCARTES, René. *Discurso do método; Meditações; Objeções e Respostas; As Paixões da Alma; Cartas*. 2.ed. São Paulo: Abril Cultural, 1979.

DIJKSTERHUIS, Ap. Think different: the merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology*, v. 87, n. 5, p. 586–598, 2004.

FREUD, Sigmund. *Vorlesungen zur Einführung in die Psychoanalyse*: Neue Folge der Vorlesungen zur Einführung in die Psychoanalyse. Frankfurt am Main: Fischer Taschenbuch Verlag, 1982.

GENNARO, Rocco J. *Consciousness and Selfconsciousness*: a defense of the higher order thought theory of consciousness. Amsterdam; Philadelphia: John Benjamins Publishing, 1996.

_____ (Org.). *Higher-order theories of consciousness*: an anthology. Amsterdam; Philadelphia: John Benjamins Publishing, 2004.

GREENWALD, Anthony G. New look III: Unconscious cognition reclaimed. *American Psychologist*, v. 47, p. 766-779, 1992.

GÜZELDERE, Güven. Introduction – Many faces of consciousness: a field guide. In: BLOCK, Ned; FLANAGAN, Owen; GÜZELDERE, Guven (Eds.). *The nature of consciousness*: philosophical debates. Cambridge, MA: MIT Press, 1997. p. 01-67.

HABERMAS, Jürgen. *Pensamento pós-metafísico.* Rio de Janeiro: Tempo Brasileiro, 1990.

KEMMERLING, Andreas. *Ideen des Ichs*: Studien zur Philosophie Descartes'. Frankfurt am Main: Vittorio Klostermann, 2005.

KIM, Jaegwon. *Philosophy of mind*. Boulder: Westview Press, 1996.

KRIEGEL, Uriah. Consciousness and Self-consciousness. *The Monist*, v. 87, n. 2, p. 182-205, 2004.

_____. *Subjective consciousness*: a self-representational theory. Oxford: Oxford University Press, 2009.

LOFTUS, Elizabeth F.; KLINGER, Mark R. Is the unconscious smart or dumb? *American Psychologist*, v. 47, p. 761-765, 1992.

PRATA, Tárik. A. A estrutura da autoconsciência na filosofia da mente de John Searle. *Veritas*, v. 62, n. 2, p. 428-452, 2017.

_____. As perspectivas de Brentano e Sartre sobre a auto-referencialidade da consciência. *Rapsódia*, v. 10, p. 42-62, 2016.

_____. Uma objeção à concepção disposicional dos fenômenos mentais inconscientes. *Principia*, v. 22, n. 3, p. 507-526, 2018.

ROSENTHAL, David. Two concepts of consciousness. *Philosophical Studies*, v. 49, p. 329-359, 1986.

ROSENTHAL, David. A theory of consciousness. In: BLOCK, Ned; FLANAGAN, Owen; GÜZELDERE, Guven (Eds.). *The nature of consciousness*: philosophical debates. Cambridge, MA: MIT Press, 1997. p. 729-753.

_____. Higher-order theories of consciousness. In: McLAUGHLIN, Brian; BECKERMANN, Ansgar (Eds.). *Oxford handbook of the philosophy of mind*. Oxford: Oxford University Press, 2008a. p. 239-52.

ROSENTHAL, David. Consciousness and its Function. *Neuropsychologia*, v. 46, p. 829-840, 2008b.

_____. Uma teoria da consciência. *Perspectiva Filosófica*, v. 44, n. 2, p. 143-178, 2017.

RYLE, Gilbert. *The concept of mind*. Harmondworth: Penguin Books, 1970.

SARTRE, Jean-Paul. *O imaginário*: psicologia fenomenológica da imaginação. São Paulo: Editora Ática, 1996.

SCHELLENBACHER, Manfred. Sigmund Freud und Franz Brentano. *E-Journal für Philosophie der Psychologie*, v. 15, p. 01-07, 2011.

SEARLE, John R. *The rediscovery of the mind*. Cambridge MA: MIT Press, 1992.

_____. *A redescoberta da mente*. São Paulo: Martins Fontes, 1997.

_____. *Mind*: a brief introduction. Oxford: Oxford University Press, 2004.

VAN GULICK, Robert. Higher-Order Global States (HOGS): an alternative higher-order model of consciousness. In: GENNARO, Rocco (Ed.). *Higher-order theories of consciousness*: an anthology. Amsterdam; Philadelphia: John Benjamins Publishing, 2004. p. 227-254.

_____. Consciência. *Investigação Filosófica*, v. E2, Artigo digital 2, 2012.

Recebido em: 08-03-2019

Aceito para publicação em: 08-07-19

ANÁLISE DA *EPÍSKEPSIS TÔN ONOMÁTON* DE ANTÍSTENES

ANALYSIS OF EPÍSKEPSIS TÔN ONOMÁTON OF ANTISTHENES

JOEDSON SILVA SANTOS[1]
Universidade Federal do Espírito Santo (UFES) – Brasil
filos_joedson@yahoo.com

RESUMO: Esta pesquisa apresenta uma análise da filosofia de Antístenes sobre a *epískepsis tôn onomáton* – investigação dos nomes. Este tema está relacionado ao problema da *orthótes onomáton*, que esteve sempre envolvida nas atividades dos sofistas. Nesse aspecto, a *orthótes* de Pródico, possivelmente um precedente inspirador para a análise antistênica dos nomes, está relacionada com a *epískepsis* de Antístenes naquilo que convergem e divergem as duas perspectivas. Tanto Pródico quanto Antístenes convergem na mesma base filosófica do princípio do *oikeîos lógos* e, consequentemente, se identificam na mesma hipótese de uma linguagem objetiva que está estritamente ligada às coisas. Mas, os dois pensadores divergem não só no problema da relação entre linguagem e realidade como também no problema da polissemia. Além do mais, expõe-se na conjectura textual o exame do pressuposto filosófico antistênico, a saber, o *oikeîos lógos*, concernente ao problema interpretativo da passagem que Aristóteles menciona Antístenes: "*oikeîos lógos hèn eph´ henós*", ou seja, enunciado próprio para cada coisa. A questão debatida é se a expressão "*hèn eph´ henós*" faz referência ao *ónoma* ou ao *lógos*.

PALAVRAS-CHAVE: Epískepsis. Orthótes. Oikeîos lógos. Pródico. Antístenes.

ABSTRACT: *This paper presents an analysis of Antisthene's philosophy on epískepsis tôn onomáton - investigation of names. This theme is related to the problem of the orthótes onomáton that was always involved in the activities of the Sophists. In this respect, the orthótes of Prodicus, possibly an inspiring precedent for the Anthistenic analysis of names, is related to the epískepsis of Antisthenes in which the two perspectives converge and diverge. Both Prodicus and Antisthenes converge on the same philosophical basis of the principle of oikeîos lógos, and consequently subscribe to the same hypothesis of an objective language, which is strictly connected to things. However, the two thinkers differ not only in the problem of the relationship between language and reality but also in the problem of polysemy. Moreover, the textual conjecture examines the philosophical Antisthenic presupposition, namely, the oikeîos lógos, concerning the interpretative problem of a passage in which Aristotle mentions Antisthenes: "oikeîos lógos hèn eph' henós", that is, a statement for each thing. The question discussed is whether the expression "hén eph' henós" refers to the ónoma or the lógos.*

KEYWORDS: *Epískepsis. Orthótes. Oikeîos lógos. Prodicus. Antisthenes.*

INTRODUÇÃO

Antístenes foi um dos seguidores de Sócrates considerado como o membro mais representativo dos discípulos após a morte do mestre. Sua importância é evidenciada nas fontes como um autor prolífico e estimado na antiguidade e, ao

[1] Mestre em Filosofia pela Universidade Federal do Espírito Santo (UFES).

mesmo tempo, foi criticado – injustamente por Timón – por escrever demais em todo tipo de assunto. Diógenes Laércio atribuiu-lhe aproximadamente setenta títulos em dez tomos com assuntos muito diversos, dos quais se destacam: crítica literária, lógica, linguagem, ética, epistemologia, ontologia, retórica e teologia. Por essa razão, esse filósofo polímata foi censurado como "charlatão universal" por Timón pela quantidade de obras escritas.[2] Reconhecido e admirado como um pensador e estilista em prosa, Antístenes desfrutou de uma excelente reputação na antiguidade (KENNEDY, 2017, p. 12-15). Ao lado de Platão e Demóstenes, foi considerado como um dos melhores expoentes do simples e puro estilo ático;[3] adjacente a Platão e Xenofonte é apontado como escritor de habilidade precisa, possuidor da "técnica de expressão"[4] e de boa reputação.[5] Apesar da notabilidade e relevância de Antístenes na era clássica e helenística, poucos estudos foram elaborados e, academicamente, parece ignorado nos dias atuais.

Este artigo tem como finalidade apresentar a filosofia lógica-linguística do filósofo ateniense, tendo como ponto de partida a análise da *epískepsis tôn onomáton*, isto é, a investigação dos nomes. Este tema é o ponto chave para entender a filosofia de Antístenes, pois direciona para vários assuntos tratados por este socrático. Esta análise do ateniense é considerada como a interpretação do *exetázein* de Sócrates; estava também associado com a crítica homérica; era tratado como o princípio para educação, também como o ponto inicial para chegar à moral e ao *oikeîos lógos*: um enunciado/nome próprio; estava também relacionado ao problema da *orthótes onomáton* que esteve sempre envolvida nas atividades dos sofistas.

Nosso trabalho pretende estudar o método de análise terminológico antistênico, a saber, *epískepsis tôn onomáton* em relação (1) aos seus aspectos gerais, (2) ao problema da *orthótes tôn onomáton*: nesta etapa, partiremos do sofista Pródico – através da metodologia de análise da correção dos nomes – como um precedente inspirador para a investigação dos nomes de Antístenes, e relacionar os dois pensadores naquilo que convergem para finalmente estabelecer as divergências das duas perspectivas. Apresentaremos (3) um exemplo da metodologia de Antístenes sobre o termo *polutropos*: neste momento, vamos averiguar o termo com base no seu radical *trópos*, analisando o seu problema polissêmico e estabelecendo os três usos específicos do nome. Por fim, sobre estas bases, trataremos do ponto final da investigação dos nomes, isto é, (4) o *oikeîos lógos*: nesta seção examinaremos se a expressão "*oikeîos lógos hèn eph´ henós*" faz referência ao *ónoma* ou ao *lógos*, e chegaremos no desfecho dos enunciados simples e compostos, ou seja, os simples quando se referem à essência das coisas

[2] Diógenes Laercio, VI.15-18 (*SSR*, V.A.41), Cf. Máralém (2014, M 792, p. 170-173) e Prince (2015, P 41A, p. 120-125).
[3] Focio, *Biblioteca*, cod. 158 (a partir de Frínico) (*SSR*, V.A.50), Cf. Máralém (2014, M 808, p. 177) e Prince (2015, P 50, p. 179-180).
[4] Longinus, *Ars Rhetorica, p. 559* (*DC 11*), Cf. Máralém (2014, M 806, p. 176) e Prince (2015, P 48, p. 177-178).
[5] Epicteto, *Disertación*, II.17.35 (*SSR*, V.A.46), Cf. Máralém (2014, M 804, p. 176) e Prince (2015, P 46, p. 175).

(*ousia*) dizem respeito ao *ónoma*, e os compostos quando fazem alusão ao *pragma* referem-se ao *lógos*.

1 Aspectos gerais da *episkepsis tôn onomáton*

Segundo o testemunho de Diógenes (*SSR*, V.A.41),[6] Antístenes foi autor de um trabalho pedagógico intitulado *"Sobre a educação ou sobre os nomes*" e um tratado erístico *"Sobre o uso dos nomes"* que constituem dois dos assuntos do tomo VII, que compõem um projeto educativo que aborda uma noção geral de educação e seus conteúdos essenciais (BRANCACCI, 2019, p. 27-28). De acordo com a concepção antistênica, os nomes podem ser agrupados em três processos: *epískepsis tôn onomáton* – investigação dos nomes – *khrêsis tôn onomáton* – uso dos nomes – e *dialegein katá géne* – distinção em classe.

A *epískepsis tôn onomáton* é testemunhada por Epicteto:

> 'Também a lógica é infrutífera'. Nós também veremos isso, mas mesmo que alguém conceda isso, o ponto anterior é suficiente, que para outras coisas a lógica é um instrumento para distinções e exames e, por assim dizer, tomar medidas e quem diz isso? Apenas Crisipo e Zeno e Cleanthes? Antístenes não diz isso? E quem escreveu que *o exame de nomes é o começo da educação?* Só Sócrates diz isso? E sobre quem Xenofonte escreve que ele *começou a partir do exame dos nomes, o que cada um significa?* (PRINCE, 2015, P 160, p. 41 – grifos acrescentados).[7]

Segundo Susan Prince (2015, p. 541-542), esse discurso de Epicteto faz parte da tese de que "a lógica é necessária", tendo como finalidade principal explicar porque os estóicos colocam a lógica em primeiro lugar em seu currículo, antes da física e da ética. Nesse aspecto, para Prince, esta passagem apoia a concepção de que "Antístenes considerou os nomes fundamentais e as unidades básicas da *τά λογικά*, que contém um sentido lógico (não retórico) no discurso completo de Epicteto". Portanto, esse testemunho parece aludir que o filósofo ateniense concebeu os nomes primários, em certo sentido, sobre a realidade, a saber, os nomes "fornecem a escala de medida para a física e a ética, e não o contrário".

Partiremos agora para uma análise central do fragmento. A máxima principal da filosofia de Antístenes: "a investigação dos nomes é o princípio da educação" (*archè paideúseos he tôn onomáton epískepsis*), é atribuída por Epicteto a Antístenes como se fosse uma citação, possivelmente da obra *"Sobre a educação ou sobre os nomes*", ou talvez pode ser também a sentença inicial, pelo motivo que "nessa obra fundamental [...] *paideia* e pesquisa sobre os nomes estavam intimamente ligadas" (BRANCACCI, 2019, p. 101, ver também PRINCE, 2015, p. 543). A expressão "a investigação do nome"[8] parece comum para Epicteto tanto

[6] Cf. Máralém (2014, M 792, p. 170-173) e Prince (2015, P 41A, p. 120-123).
[7] Cf. Mársico (2014, M 979, p. 259).
[8] A doutrina educativa dos nomes de Antístenes não está apenas na passagem de Epicteto, em um contexto lógico-pedagógico, ela se encontra embutida na crítica literária de Homero, em especial

em Antístenes como em Sócrates, mas nenhuma literatura socrática existente faz essa atribuição a Sócrates, o que se aproxima aparentemente é o que Xenofonte em "Memoráveis" diz:

> Sócrates acreditava que aquele que sabe o que é cada coisa pode também explicá-lo a outros; agora, os que não sabem nada, esses – dizia ele – seria de esperar que se enganassem e enganassem os outros. Por essa razão, dizia que nunca deixava de examinar, com aqueles que o acompanhavam, a essência de cada coisa. Seria difícil explicar como é que construía todas estas definições, mas acho que o que vou contar é suficiente para demonstrar qual a *sua metodologia de investigação.* (XENOFONTE, 2009, IV.6.1, p. 268 – grifos acrescentados).

Segundo Prince (2015), o contexto de Xenofonte mostra que ele está fazendo menção de definição – por exemplo de piedade – e não de uma investigação de nomes, mas do que existe ou das coisas; diferente de Epicteto ele não usa o termo "significar". Nesse sentido, se, e somente se, a referência de Epicteto for a obra "*Memoráveis*", podemos inferir que o mesmo se equivocou na sua interpretação do texto, entretanto, por outro lado, pode-se dizer, como faz Prince, que "Epicteto pode estar citando não *Mem.* IV.6.1, mas um texto semelhante"; ela ainda acrescenta, "Se a comparação de 'Sócrates' e Antístenes de Epicteto é precisa, isso poderia ser porque Xenofonte usou a linguagem de Antístenes para descrever o procedimento educacional de Sócrates" (PRINCE, 2015, p. 543; BRANCACCI, 2019, p. 139-159).

O vocábulo "princípio da educação" (*archè paideúseos*) é interpretado por Prince (2015) como um passo geral na educação como um todo, e a organização do tomo VII do catálogo de Diógenes implica também uma sequência geral e, quem sabe, uma necessidade de entender a natureza geral dos nomes antes de avançar na aprendizagem dos livros consecutivos.[9] Para Brancacci (2019) não é puramente uma dimensão pedagógica, porque em torno deste problema é tematizada a questão metodológica de captação do real. Nas palavras do autor, "se o exame dos nomes é de fato tomado como elemento da educação (παιδεύσις), ele é, também, propriamente o princípio (ἀρχὴ) de sua atualização e o ponto de partida de seu processo de aquisição (BRANCACCI, 2019, p. 102). Assim sendo, a passagem de Epicteto pode ser considerada tanto como uma máxima para o sistema educacional como para apreensão da realidade, e também para o "princípio ou fundamento da formação intelectual" (MÁRSICO, 2014, nota 205, p. 259).

A *epískepsis* é considerada como uma interpretação de Antístenes da *exetázein* de Sócrates, assimilada a uma análise de termos, tendo como finalidade

nos fragmentos de Porfírio (*SSR*, V.A.187) sobre *polýtropos* (abordaremos mais à frente) e (*SSR*, V.A.189) sobre Cíclopes; bem como nos discursos de Ajax (*SSR*, V.A.53) e de Odisseu (*SSR*, V.A.54).

[9] *Sobre o uso dos nomes* (tratado erístico), *Sobre a pergunta e a resposta*, *Sobre a opinião e o conhecimento* (em quatro livros), *Sobre o morrer*, *Sobre a vida e a morte*, *Sobre o assunto do Hades*, *Sobre a natureza* (em dois livros), *Pergunta sobre a natureza primeira*, *Pergunta sobre a natureza segunda*, *Opiniões ou erístico*, *Sobre o aprender* (problemas) (DIÓGENES LAERCIO, VI.15-8 (*SSR*, V.A.41)).

a determinação de cada um deles, fornecendo uma definição na conclusão do processo dialético (BRANCACCI, 2005, p. 12). O termo grego *exétasis* ou *exatázein* significa "submeter a exame", "interrogar", "pôr à prova", isto é, um método dialético associado tanto por Platão quanto por Xenofonte como uma metodologia de Sócrates em um processo de verificação ou exame de nomes. Brancacci (2019) identifica esse método socrático testemunhado por Xenofonte com o de análise dos nomes de Antístenes, possivelmente porque Sócrates agrupava os termos *dialégesthai,*[10] *syneînai*[11] e o *exetázein* ao método dialético de perguntas e respostas em sua relação com o sábio. Nesse sentido, o *sophós* é aquele que é apto para realizar o método de exame (*exetázein*) para comunicar o saber adquirido mediante o diálogo ou discussão (*dialégesthai*) de um tipo de assunto para desenvolver um bom diálogo entre os homens (*anthrṓpois syneînai*). Nas palavras de Brancacci,

> O método da divisão em classe constitui a condição da capacidade de discutir e conversar com os homens (**διαλέγεσθαι καὶ ἀνθρώποις συνεῖναι**), porque ele determina a aquisição de conteúdos de saber que os sábios, através de seu ensinamento, transmitirão aos homens. Disso deduz-se que, graças à transformação do conceito do **ἐξετάζειν**, entendido por Antístenes como um exame dos nomes, a noção de *dialegesthai* torna-se autônoma. Essa transformação explica também a atribuição aos sábios de uma habilidade ao mesmo tempo dialética, já que ligada à aptidão para efetuar corretamente o exame dos nomes, e retórica, dado que a virtude de Odisseu, o *polutropos*, designa uma capacidade de instruir os homens, encontrando o tipo de discurso apropriado a cada um. (BRANCACCI, 2019, p. 173).

Brancacci (2005, p. 12) ressalta que a concepção dialética de Antístenes era diferente da que Platão atribuiu a Sócrates. Para o Sócrates platônico a resposta à questão da definição era o ponto de partida para *exetázein,* que considera como uma "situação dialógica concreta na qual o princípio ético fundamental foi ativado, isto é, o princípio em que a *dialegesthai* era o bem supremo"; diferentemente, Antístenes compreende a definição da qualidade peculiar, a saber, o *oikeîos lógos*, como ponto de chegada da *epískepsis tôn onomáton*. Deste modo, para o Sócrates platônico, o *exetázein* é identificado com o *dialegesthai*, que pareceu a Antístenes ser um verdadeiro método, porém não equivalente ao bem supremo (*megiston agathon*), mas sim o princípio da educação (*archè paideúseos*).

A investigação dos nomes, na sua finalidade de educar, tem como consequência lógica a utilização correta dos nomes. Antístenes escreveu um livro intitulado "*Sobre o uso dos nomes*" (*SSR*, V.A.41), tendo a noção de *khrêsis tôn*

[10] O termo *dialegesthai* (dialogar) tem a mesma raiz de *dialegein* (catalogar), cujo verbo, *legein*, pode ser entendido como "reunir", "contar" e, consequentemente, pode ser compreendido também como "enumerar", "calcular", "discutir". *Dialegesthai*, nas palavras de Prince, é um termo positivo para Antístenes, porque é um aspecto da habilidade do sábio retórico; Xenofonte chama de método socrático, porém esse termo não está se referindo a dialogar, em Mem. IV, 5.12, mas de ordenação ou classificação (*dialegein*). Cf. Prince (2015, p. 148); Brancancci (1990, p. 149-152) e Xenofonte (IV.6.1, 2009, nota 179, p. 268).

[11] O termo *syneînai* tem como significado conversar entre si ou conversar uns com os outros.

onomáton como o segundo processo do núcleo conceitual da proposta antistênica, a saber, o método de análise terminológico. A expressão *khrêsis*[12] (uso) pode estar associada a *orthótes khrêsis* (uso correto) e *katákhrêsis* (uso incorreto). Além do mais, esse método de análise está estritamente vinculado com o saber prático, pois o termo *khrêsis* tem implicações éticas em Antístenes. Portanto, o método de análise onomástico constituído pelos processos de investigações (*epískepsis*), uso correto (*orthótes khrêsis*) dos nomes e por sua distinção de classe (*dialegein katá géne*)[13] permite indicar ou revelar o nome para cada coisa, isto é, o *oikeîos lógos.*

2 *Orthótes* e *epískepsis tôn onomáton*: Pródico e Antístenes

O método de análise de Antístenes parece estar relacionado ao problema da *orthótes onomáton*,[14] que esteve sempre envolvida nas atividades dos sofistas. O *orthós lógos* dos sofistas, em especial do *ónoma*, estava desenvolvido sobre a problemática do binômio *nómos/phýsis* – assim como da disjunção e inerência entre linguagem e ser – por meio da antiga disputa sobre a origem da linguagem: convencionalismo/naturalismo, a saber, se há entre o nome e o ente uma coincidência natural ou uma identificação convencional. O conceito de *orthótes* em Protágoras pressupõe a doutrina da antilogia, ou melhor, dos dois *lógoi* opostos para cada *prágma*. Em outras palavras, para cada coisa há dois *lógoi* opostos, dos quais um é mais forte ou superior que o outro. Essa perspectiva entra em choque com as concepções de pensadores naturalistas tais como Pródico e Antístenes, que sustentaram o *eikeîos lógos*, isto é, um enunciado para cada coisa, a fim de estabelecer uma relação única entre nome e coisa (*ónoma-prágma*).

A correção dos nomes em Pródico tem como base o *oikeîos lógos* e o estudo das palavras pela qual distinguia o sentido das classes das palavras, por meio do método dierético (*diáiresis*) ou distinção, que consistia na "determinação do significado de uma palavra e, acima de tudo, de sua diferença com outra homônima ou sinônima que exija comparar suas formas globais ou parciais e, portanto, averiguar sua etimologia" (DOMÍNGUEZ, 2002, p. 48). Em outros termos, esse método versava o exame sistemático das diferenças e oposições entre palavras aparentemente sinônimas (MELERO BELLIDO, 1996, nota 26, p. 247), com o propósito de restringir o uso da linguagem ao que descreve a coisa, e que em sua própria estrutura manifeste ou indique ou revele também a estrutura da realidade.

[12] A questão do uso é marcante na filosofia de Antístenes, além da sua preocupação com o uso dos nomes, ele também escreveu um livro – mencionado no tomo IX (*SSR*, V.A.41) – "Sobre o uso do vinho" e abordou – segundo o testemunho de Epicteto (*SSR*, V.B.22) – sobre *khrêsis phantasion*, ou seja, uso da aparência ou da imagem.

[13] Para uma análise pormenorizada dessa expressão ver Máralém (2005a, p. 88-91; 2005b, p. 123-125).

[14] Platão atribuiu essa atividade aos sofistas em geral (CRÁTILO, 291b.), a Protágoras (CRÁTILO 291c; FEDRO 267c) e a Pródico (CRÁTILO 384b; EUTIDEMO 277e). Platão apresenta também no livro *Crátilo ou Correção dos Nomes* dois interlocutores de Sócrates, Hermógenes e Crátilo, que estavam mergulhados nessa problemática. Enquanto o primeiro assumia uma posição convencionalista da linguagem, o segundo admitia um naturalismo que, em certo sentido, estava mais próximo de Pródico e Antístenes.

O *oikeîos lógos* e, consequentemente, a explicação correta da estrutura da realidade de Pródico estão associados com a noção da adequação dos nomes, atribuído por Platão no Crátilo (384b), no Eutidemo (277e-278b) e no Cármides (163d). Para Pródico, um nome só tem sentido se for nome de uma coisa, se um nome não é nome de alguma coisa não tem significação. Nesse sentido, Kerferd (2003, p. 124) fala que "cada segmento da realidade pertence a um *lógos*, e cada *lógos* corresponde exatamente a um segmento distinto da realidade". Logo, Pródico utiliza do método *diáiresis* orientado pelo princípio do *oikeîos lógos* para buscar o uso unívoco da linguagem, propondo vincular cada nome a uma coisa.

Provavelmente, essa metodologia de análise da correção dos nomes tenha notórias influências em Antístenes. Ainda que não haja evidências diretas para as opiniões de Antístenes sobre sinônimos, esta concepção pode ser inferida por seu princípio do *oikeîos lógos* e de seu paradoxo da impossibilidade de contradizer (*ouk éstin antilégein*) (PRINCE, 2015, p. 55); além do mais, pode-se encontrar um projeto próximo à análise de homônimos no testemunho de Porfírio, onde Antístenes distingue os vários sentidos de uma palavra (SSR, V.A.187).

Segundo Máralém (2005a, p. 75-76; 2005b, p. 112-113) e Brancacci (2019, p. 69), o *orthótes* de Pródico foi um precedente inspirador para a teoria de Antístenes, esta aparece nas fontes como *epískepsis tôn onomáton* ou *khrêsis tôn onomáton*. Esse método de análise testemunhado por Epicteto (*Disertaciones*, I.17.10-12 (*SSR*, V.A.160)), isto é, "a investigação dos nomes é o princípio da educação", parece compartilhar o campo de nomes com Pródico. Esta afirmação pode ser testificada pelo relato de Platão (EUTIDEMO 277e-278a) quando se refere à fala de Pródico no diálogo Eutidemo: – "é necessário aprender sobre a correção dos nomes". Não só nessas duas máximas, que em certo grau parecem equivalentes, encontramos uma possível influência de Pródico em Antístenes. Além de seus interesses comuns aos nomes, eles compartilham um interesse na história de Héracles[15] e, possivelmente, em uma teoria sobre o uso (*khrêsis*) como base para a ética[16] (PRINCE, 2015, p. 55). Esses pensadores também convergem na mesma base filosófica do princípio do *oikeîos lógos* (um enunciado para cada coisa) e, consequentemente, se identificam na mesma hipótese de uma linguagem objetiva que está estritamente ligada às coisas.

Apesar de algumas convergências, deve-se clarificar que a postura de Antístenes é mais radical. Sendo assim, é possível estabelecer divergências entre estes dois pensadores. Para Máralém (2005a, p. 77; 2005b, p. 113), uma das diferenças que há entre as duas abordagens está na substituição dos termos *orthótes* pelo *epískepsis*. Contrariamente, discorda-se que haja uma substituição dos termos, acredita-se que haja uma ampliação dele. Máralém tenta justificar o fato da *orthótes* de Pródico deixar dúvida sobre a relação entre linguagem e realidade. Segundo ela,

> estritamente falando, depois da noção de **ὀρθότης**, na medida em que este termo implica a 'correção' como um estado, mas também

[15] Conferir o testemunho em *SSR*, V.A.92-99 e *SSR*, V.A.207, IV.A.224.

[16] Ver nota sobre o termo **τὴν τοῦ λόγου χρῆσιν** sobre o testemunho de Porfírio (SSR, V.A. 187) em Prince (2015, p. 606-607).

> como um processo, esconde-se a ideia de que a correlação entre linguagem e realidade, se bem existe, nem sempre é clara e efetiva, Portanto, a tarefa do **ὀρθότης** é verificar, isto é, corroborar a correção, ou restituí-la, ou seja, efetuar a correção. (MÁRSICO, 2005a, p. 77).

Entende-se que a *orthótes* de Pródico é diferente pela mesma razão que Márs ico (2005a) a justifica, no entanto, isso não implica dizer que há substituição dos termos. Essa tomada de decisão da intérprete pode estar relacionada ao fato de que não aparece o termo *orthótes tôn onomáton* em Antístenes. Compreende-se que os termos *epískepsis* e *khrêsis* não estão dissociados do *orthótes*, mas são uma ampliação da análise, já que o *orthótes* é restrito somente à correção; contudo, para que haja correção é preciso investigar e usar adequadamente o nome. De fato, o *orthótes* de Pródico deixa obscuro a relação entre linguagem e realidade, mas é possível perceber que sua análise gira em torno dos termos – utilizando as palavras da professora – "verificar, corroborar, restituir e efetuar" a correção. Esses termos não são substitutos da *orthótes*, mas sim fazem parte da metodologia da mesma; assim como *epískepsis* e *khrêsis* são análises semânticas dos nomes para diferenciar os vários significados próximos a um termo e determinar o uso próprio ou adequado do termo para resolver o problema do *orthótes tôn onomáton*. Logo, entende-se que não há substituição de um termo por outro e sim uma ampliação da análise.

A segunda divergência, já supracitada, se baseia no problema da relação entre linguagem e realidade. Enquanto para Pródico essa correlação não é sempre clara e efetiva, para o socrático Antístenes,

> o pressuposto dado é que a correlação sempre existe e, nos casos escuros, é necessário simplesmente realizar uma análise - **ἐπίσκεψις** - do termo que permitirá mostrar que a relação entre linguagem e realidade é perfeita, e que cada coisa corresponde um nome; isto é, que cada coisa tem seus oikeîos lógos, seu próprio e único nome. (MÁRSICO, 2005a, p. 77; Cf. 2005b, p. 113).

Tanto Brancacci (2019) como Márs ico (2005a) apresentam um exemplo da atividade de Pródico relatado por Platão no Protágoras (337a-c) cujo procedimento estava orientado a averiguar o correto significado de uma palavra para apontar a exata adequação entre nome e coisa. Apesar de Pródico crer no significado objetivo dos termos, parece que aceita algum tipo de formulação que suponha a convencionalidade da linguagem (BRANCACCI, 1990, p. 63; MÁRSICO, 2005a, p. 76; 2005b, p. 113). Esta hipótese da possível aceitação de um tipo de convencionalismo de Pródico pode ser justificada não só pela obscuridade da relação entre linguagem e realidade, se é que existe, mas também pelo problema da polissemia.

Segundo Márs ico (2005a, p. 78), em Pródico, averiguado um caso de polissemia, era preciso corrigi-lo tendo em conta o princípio do *oikeîos lógos*, a saber, de que cada coisa deve corresponder um nome, a fim de que, quando produzia polissemia se estava frente a um mal-uso da linguagem, pelo que deveria

restituir ou verificar e, por fim, efetuar a correção. Para Brancacci (2019, p. 69), Pródico interpretava a polissemia como mera oscilação do *onomazein*, que necessitava, portanto, de correção: através da exigência normativa de fixar um nome para cada coisa que lhe correspondesse. O uso próprio do nome à coisa estava aplicado a uma revisão da nomenclatura, dirigida a excluir a possibilidade de uma efetiva multiplicidade de significados das palavras. Nesse sentido, o *orthótes* de Pródico estava vinculado com os demais sofistas no que concerne à necessidade de introduzir modificações na estrutura da linguagem para estabelecer o *orthótes tôn onomáton* (MÁRSICO, 2014, p. 38).

Nesse mesmo aspecto, pode-se delinear a terceira diferença entre esses dois pensadores. Para Pródico, a polissemia era um problema ou um mal-uso da linguagem e era preciso uma revisão ou modificação na estrutura da linguagem, assim como a exclusão da possibilidade de vários significados.

Em Antístenes, entretanto, a polissemia não era um problema, mas um dado linguístico que não necessitava de uma alteração da estrutura da linguagem, mas sim de uma investigação dos nomes.

> Para Antístenes, por outro lado, a polissemia era um fato linguístico de que não havia necessidade de negar nem requerer de uma conduta ativa em prol de sua correção: era necessário apenas determinar com clareza sua esfera de aplicação, isto é, estudar o *khrêsis tôn onomáton*, para que a legalidade efetivamente presente na língua se manifeste. Assim, também no caso de Antístenes se chegava à correlação de 'um nome para cada coisa', mas não por alteração do dado linguístico, mas por simples estudo lexical. (MÁRSICO, 2005a, p. 78).

Além do mais, a multiplicidade de significados era mantida aberta, determinando assim, com claridade a legítima esfera de uso ou aplicação de cada um deles (BRANCACCI, 2019, p. 70).

3 Polutropos

Um exemplo da metodologia de análise de Antístenes está testemunhado por Porfírio. Nesse testemunho, pode-se averiguar claramente o mecanismo da *epískepsis* e *khrêsis* concernente ao nome *polutropos*.[17] Esse epíteto fora atribuído a Odisseu por Homero e era entendido no sentido de caráter (*tropos*) variável (*polu*) que estava associado a um homem mentiroso (ou malvado). De acordo com esse entendimento, Homero parece ter um desprezo pelo herói ou o denuncia pela sua astúcia enganosa. No testemunho de Porfírio, Antístenes se propõe refutar essa interpretação por meio do método da *epískepsis tôn onomáton*, verificando o termo *polutropos*, com base no seu radical *trópos*, demarca o problema polissêmico de *trópos* delimitando os três usos (*khrêsis*) específicos do nome em

[17] Para uma análise pormenorizada do termo *polýtropos* cf. M. Luzzatto, "*Dialettica o retorica? La polytropia di Odisseo da Antistene a Porfirio*", en *Elenchos* 17, Napoli, 1996, pp. 275-358.

seu contexto ou sentido apropriado para mostrar que a atribuição do termo *polutropos* a Odisseu não está associada com o sentido ético.

Em seu testemunho, Porfírio[18] diz que

> Antístenes afirma que Homero não elogia nem critica a Odisseu quando o chama de "multifacetado" *(polutropos)*. [...] Então, ao analisar, Antístenes disse: E depois o que? Acaso é mal Odisseu porque foi chamado de *polutropos*, e não é possível pensar que Homero o chamou assim porque era sábio? Acaso *trópos* não significa em um aspecto o caráter e em outro significa o uso do discurso? Pois *eútropos* é o varão que tem o caráter voltado para o bem, e *trópoi* dos discursos que são de vários estilos. E Homero também usa o termo *trópos* em relação a voz e as melodias variadas, como no caso do rouxinol, que frequentemente gorjeando expande sons variadamente modulados. E se os sábios são hábeis para falar, sabem dizer o mesmo conceito de muitos modos e conhecendo muitas maneiras de argumentar sobre os mesmos seriam *polutropos*. Por isso disse Homero que Odisseu, por ser sábio, é *polutropos*, porque sabia conviver com os homens de muitos modos. (MÁRSICO, 2014, M 1011, p. 272-273).

Antístenes utiliza-se de um procedimento de análise léxica para descobrir o sentido do termo *polutropos* através dos sentidos de *tropos*, que se identificam em três âmbitos: o primeiro âmbito é o ético – com relação ao caráter –, o segundo é o retórico – multiplicidade de modos discursivos – e, finalmente, o terceiro no campo da música – variação de voz e melodia. Segundo Claudia Mársico (2005, p. 81), o desafio desse procedimento é "explicar os três usos sem que a noção perca especificidade". No primeiro âmbito, isto é, o ético, a análise se dá pela etimologia e exige a inclusão dos termos *trépo* e *eútropos*: *trópos* se unifica com *trépo* "girar", "dar volta", e a palavra *eútropos* é o que se orienta ao bem. Já no âmbito retórico, o exame se dá pela semântica e se sustenta na relação de significado entre os termos *trépo* e *plásso* ("modelar", "forjar"). No terceiro caso, refere-se aos estilos de sons. Assim como no segundo âmbito, esse reintroduz a categoria de multiplicidade tal como aparece em *polutropos* como variedade de sons e melodias. Portanto, *polu*, unido com a noção de *tropos*, foi associada à ideia de multiplicidade sem que se implique o sentido negativo. Nesta interpretação, Homero associa o termo *polutropos* não no sentido ético, com menosprezo ou crítica ao herói, mas no âmbito retórico, associado à multiplicidade de modos discursivos. Portanto, Homero faz referência a Odisseu no sentido de um *sophós* (sábio), um homem com habilidade multifacetada de modos discursivos. Logo, só um sábio pode dar conta da multiplicidade do real, porque poderá entender a trama do existente e designar a cada coisa o nome que lhe é próprio, ou seja, o *oikeîos lógos*. De acordo com este princípio, a linguagem é única, ou melhor, sempre adequada à coisa, o que torna múltipla a necessidade de gerar diversos discursos que só são possíveis porque os que proferem – em especial um *sophós* – sabem precisamente que só há um *lógos* para cada coisa.

[18] Porfirio, *Escolio a* Odisea, I.1 (*SSR*, V.A.187) cf. em Prince (2015, P 187, p. 591-594).

4 *Oikeîos Lógos*

Segundo Diógenes Laercio,[19] Antístenes definiu que o *lógos* "é o que mostra o que era e o que é" (*ho tò tí ên è ésti delôn*). Máralém (2014) nos informa que o termo *delôn* (que mostra) faz da linguagem um elemento que permite revelar a natureza do real. Portanto,

> a adoção do verbo *deloûn*, "mostrar", associada à revelação e evidência que surge da sinalização, declara que a linguagem não possui carências e inconveniências estruturais como as postuladas pelos megáricos (o real é um, a linguagem é múltipla, de modo que não serve para mostrar o uno), nem está afetada por insuficiências de transmissão, como no sistema gorgiano (se fosse pensado, não poderia ser transmitido sem perda de sentido. (MÁRSICO, 2014, p. 33).

Brancacci (2019, p. 239-242) entende que essa definição de *lógos*, testemunhada por Diógenes, tem a mesma significação do termo que aparece no princípio filosófico antistênico, a saber, *oikeîos lógos*. Pois, o que mostra que algo era ou é um *oikeîos lógos*. O problema do *lógos*, neste princípio, é saber se ele refere somente ao *ónoma* ou ao *lógos*. Mas antes de tratar desse problema passamos para a definição do *oikeîos*. Esse termo, em seu sentido geral, significa "próprio", "privado" ou "único", por isso o princípio antistênico é interpretado como um nome ou um enunciado para cada coisa. O problema está nas interpretações que os comentadores fazem na passagem em que Aristóteles menciona Antístenes. Segundo Aristóteles,[20] "Antístenes ingenuamente acreditava que nada é dito com relevância, exceto por meio do enunciado próprio [*oikeîos lógos*], um para cada coisa [*hèn eph´ henós*]. Daí surgiu que não é possível contradizer [*antilégein*], e nem mesmo é possível mentir [*pseudeîn*]". O problema interpretativo dessa passagem gira em torno da expressão "*oikeîos lógos hèn eph´ henós*", ou seja, enunciado próprio para cada coisa. A questão debatida entre os comentadores é se a expressão "*hèn eph´ henós*" faz referência ao *ónoma* ou ao *lógos*.

Para Cordero (2008, p. 123; 2001, p. 331-332), na passagem de Aristóteles, Antístenes não sustenta um *μακρός λόγος*, mas reduz o *lógos* ao *ónoma*. Segundo ele, os intérpretes deduzem equivocadamente a frase "um para cada coisa", como se ela se referisse a "um *lógos* (*hèn*) para cada coisa (*eph´ henós*)", mas essa interpretação não pode ser aceita, porque *hèn* é neutro, portanto, não pode fazer referência a *lógos*, por ser esse termo masculino. Pois, se fizesse alusão a *lógos*, deveria ser lido em *heîs* e não *hén*. Logo, na concepção de Cordero (2008), a interpretação correta é "um *ónoma*, neutro, para cada coisa".

Essa interpretação, baseada em critérios gramaticais, dá a entender que em Antístenes há uma relação natural biunívoca entre *ónoma* e *prágma* (nome e coisa), que o *oikeîos lógos* seja uma suposição da identidade entre nome e coisa, idêntico

[19] Diógenes Laercio, VI.3 (*SSR*, V.A.151), Cf. Máralém (2014, M 958, p. 244) e Prince (2015, P 151A, p. 473).
[20] Aristóteles, *Metafísica*, V.29.1024b26 (*SSR*, V.A.152), Cf. em Máralém (2014, M 960, p. 245-248) e Prince (2015, P 152A, p. 485-486).

a uma tautologia do tipo "A é A". Por conseguinte, o *lógos*, isto é, "o que mostra o que era ou o que é", não faz referência a um enunciado, mas a um nome, aquele que melhor mostra o que era (por exemplo o termo dinossauro) e o que é (planeta Mercúrio) sem nenhuma informação adicional.

Contrariamente, para Brancacci (2019, p. 265), a objeção gramatical da interpretação supracitada não é determinante, porquanto a expressão "*hèn eph´ henós*" pode ser entendida como uma frase geral em função apositiva em relação ao *oikeîos lógos*. Em outros termos, Brancacci (2019) justifica a relação entre a oração "*hèn eph´ henós*" com a locução nominal *oikeîos lógos*, também em critérios gramaticais. A frase, "*hèn eph´ henós*", é uma oração subordinada substantiva apositiva exercendo a função de aposto, ou seja, ela esclarece, explica um termo anterior, a saber, *oikeîos lógos*. Portanto, Brancacci (2019), em consonância com a interpretação de Alejandro de Afrodisia,[21] entende que a correlação se dá entre *lógos* próprio e *prágma* (enunciado e coisa). Brancacci (2019) ainda observa que os que fazem inferência da tautologia em Antístenes não leva em consideração o que Aristóteles diz quando faz menção de Antístenes. Segundo o filósofo estagirita,[22] "o enunciado de cada coisa é, como um, o da essência, mas também é múltiplo, pois, de algum modo, é o mesmo algo e algo afetado de certa maneira, por exemplo, Sócrates e Sócrates músico (e o enunciado falso é simplesmente enunciação de nada)". Nesse sentido, o *oikeîos lógos* pode ser uno, quando se refere à essência da coisa (*ousia*), e pode ser múltiplo, quando se refere a *prágma*. Nesse aspecto, deduzimos que a leitura tautológica de Antístenes só é válida quando se refere ao nome ou a enunciados simples (micro *lógos*), mas não a enunciados compostos ou complexos (macro *lógos*).

As duas interpretações – a de Cordero (2008) e a de Brancacci (2019) – aparentemente são diferentes, mas não são excludentes, pois uma completa a outra. Ousadamente, inferimos que o *oikeîos lógos* se refere tanto ao *ónoma* quanto ao *lógos*, tanto ao uno quanto ao múltiplo. Apesar dos dois comentadores utilizarem corretamente os critérios gramaticais, uma interpretação não pode refutar a outra. Todavia, é necessário relacionar as duas para entender a filosofia de Antístenes.

O *lógos* relacionado com o *oikeîos* pode se referir ao *ónoma*, pois em Antístenes predicar é "dar nome às coisas" (DINUCCI, 1999, p. 108), consequentemente, temos um enunciado simples ou denominativo. Um exemplo de um enunciado designativo é "este é Sócrates", "este" refere à *prágma* e "Sócrates" alude ao *ónoma* da *ousia*. Em outros termos, "Esta coisa (*prágma*) é o nome (*ónoma*) da coisa (*ousia*)" (DINUCCI, 1999, p. 109). O *ónoma* está em correlato unívoco com as coisas, está essencialmente unido ao ente. Além do mais, os nomes são imitações vocais das coisas e toda atribuição às coisas que não seja correta não é nome, mas meros sons sem sentidos.

[21] Alejandro, *Sobre la Metafísica de Aristóteles*, 434.25-435.20 (*SSR*, V.A.152), Cf. Máralém (2014, M 961, p. 249) e Prince (2015, P 152B, p. 498-499).
[22] Aristóteles, *Metafísica*, V.29.1024b26 (*SSR*, V.A.152), Cf. Máralém (2014, M 960, p. 245-248) e Prince (2015, P 152A, p. 485-486).

Segundo Aristóteles,[23] Antístenes pensava que não era possível definir "o que é" (*ti esti*), pois a definição é um enunciado largo (macro *lógos*). De acordo com esse testemunho, é impossível definir a essência em Antístenes, ou seja, os objetos simples não podem ser definidos; o nome exato é aquele das coisas que não podem ser definidos, mas podem ser descritos – como é (*poîón esti*). Já os objetos compostos podem ser definidos, isto é, recebem um *lógos*. *Lógos* é um composto de nomes dos elementos que compõem a coisa. "As coisas são tão somente uma combinação de elementos simples, uma definição nada mais é que uma enumeração dos nomes destes elementos simples que são indefiníveis" (DINUCCI, 1999, p. 114). Um exemplo citado por Dinucci (1999, p. 113) é do testemunho de Pseudo-Alexandre, que se "consideramos o nome 'homem', podemos defini-lo como animal mortal racional, obtendo um *lógos*, ou fórmula longa, composto de *onomata*, que se referem aos elementos que compõem o homem enquanto *prágma*".

Nesse sentido, estabelecemos o enunciado composto ou complexo – um *lógos* ou um nome composto por mais de uma palavra –, a saber, "Sócrates é homem-filósofo", em outros termos, a *ousia* de C (*prágma*) é X (*lógos*) equivale a dizer que a natureza da coisa – "Sócrates" – é um *lógos* – "homem-filósofo". Por conseguinte, o *ónoma* (nome) e o *lógos* (encadeamento de nomes) se referem à linguagem que expressa o pensamento sobre as coisas, a qual não só permite mostrar a *prágma* e a *ousia*, mas funciona como manifestação unívoca da estrutura da realidade.

Considerações finais

A análise da *epískepsis tôn onomáton* é o ponto inicial que desencadeia a compreensão da filosofia lógica-linguística, como também ético-pedagógica, de Antístenes. Este filósofo, pouco conhecido na área acadêmica nos dias atuais, desenvolveu uma metodologia de análise terminológica que é considerada como princípio da formação intelectual. Este processo onomástico do prolífico autor que norteia nossas reflexões serve como propedêutica para aqueles que queiram conhecer um campo de conhecimento pouco desbravado pelos estudantes brasileiros.

Seu mecanismo de análise estava associado com a problemática da correção dos nomes; possivelmente, a metodologia de Pródico foi um meio que inspirou o socrático a desenvolver sua filosofia. Deste modo, esses dois pensadores, nas suas reflexões filosóficas, compartilharam o mesmo interesse pela nomeação apropriada a cada coisa, como na crença de uma linguagem objetiva estritamente conexa às coisas. Entretanto, ambos divergem no tocante à relação linguagem e realidade, como também no problema polissêmico dos nomes.

Antístenes utiliza o procedimento da *epískepsis*, *khrêsis* e *dialegathai* como instrumento de investigação para analisar o termo *polutropos*. E,

[23] Aristóteles, *Metafísica*, VIII.3.1043b4-32 (*SSR*, V.A.150), Cf. Máralém (2014, M 956, p. 238-241) e Prince (2015, P 150A, p. 445-447).

consequentemente, refutar uma concepção errônea de que Homero tinha o desprezo pelo herói ao aplicar este termo no sentido de caráter. O socrático faz uma análise léxica do vocábulo, distingui-o em três classes – ético, retórico e musical – para chegar ao uso apropriado. Nesse sentido, entende que Homero faz referência ao herói não no sentido ético, mas no sentido retórico, ou seja, de um *sophós* com habilidade multifacetada de modos discursivos, que tem a sabedoria de designar a cada coisa o nome que lhe é próprio, a saber, o *oikêios lógos*.

A *epískepsis* é o ponto inicial para se chegar ao *oikêios lógos*, este é o princípio filosófico do ateniense, que tem como pano de fundo, o uso unívoco da linguagem. O problema que contorna esse princípio é se a expressão mencionada por Aristóteles, "*oikeîos lógos hèn eph´ henós*", faz referência ao *ónoma* ou ao *lógos*. Constatamos que, para se entender a filosofia de Antístenes, é necessário relacionar os dois, nome/logos, pois quando faz referência ao enunciado simples deve ser aplicado o *ónoma*, mas quando é utilizado um enunciado composto deve-se aplicar o *lógos* (encadeamento de nomes). Portanto, o *oikeîos lógos* se refere tanto o *ónoma* quanto ao *lógos*, tanto ao uno quanto ao múltiplo.

Referências

BRANCACCI, Aldo. Episteme and Phronesis in Antísthenes. *Méthexis*, v. 18, p. 07-28, 2005.

_____. *Oikeios logos*: linguagem, dialética e lógica em Antístenes. Tradução de Joseane Prezotto e Simone Petry. Rio de Janeiro: Ed. PUC-Rio; São Paulo: Edições Loyola, 2019.

CORDERO, Néstor. Antístenes: un testigo directo de la teoría platónica de las Formas. *Revista de Filosofía de la Universidad de Costa Rica*, v. XLVI, n. 117-118, p. 119-128, 2008.

_____. L'interprétation anthisthénienne de la notion platonicienne de "forme" (eidos, idea). In: FATTAL, Michel (Ed.). *La philosophie de Platon*. Paris: L'Harmattan, 2001. p. 323-343.

DINUCCI, Aldo L. Lógica e teoria da linguagem em Antístenes. *O Que nos Faz Pensar*, v. 13, p. 105-118, 1999.

DOMÍNGUEZ, Atilano. *Cratilo o del lenguaje*. Ed. y trad. del griego de Atilano Domínguez. Clásicos de la Cultura. Madrid: Trotta, 2002.

GIANNANTONI, Gabriele. *Socratis et Socraticorum Reliquiae*. 4 vols. Napoli: Bibliopolis, 1990.

KENNEDY, William J. *'Anthisthenes' Literary Fragments:* Edited with introduction, translations, and commentary. Thesis (Doctor of Philosophy) – Faculty of Arts University of Sydney, 2017.

KERFERD, George B. *O movimento sofista*. São Paulo: Loyola, 2003.

MÁRSICO, Claudia. Antístenes y la prehistoria de la noción de campo semântico. *Nova Tellus*, v. 23, n. 2, p. 70-99, 2005a.

MÁRSICO, Claudia. Argumentar por caminos extremos: II) La necesidad de pensar lo que es. Antístenes y la fundamentación semántica de la verdad como adecuación. In: CASTELLO, L.; MÁRSICO, C. (Eds.). *¿Cómo decir lo real?* El lenguaje como problema entre los griegos. Buenos Aires: GEA, 2005b. p. 109-132.

_____. *Filósofos socráticos*. *Testimonios* y *fragmentos II*. Antístenes, Fedón, Esquines y Simón. Buenos Aires: Editorial Losada, 2014.

MELERO BELLIDO, Antonio. *Sofistas*: testimonios y fragmentos. Madrid: Gredos, 1996.

PLATÃO. *Eutidemo*. Texto estabelecido e anotado por John Burnet. Tradução, apresentação e notas de Maura Iglésias. Edição bilíngue grego-português. Rio de Janeiro: Ed. PUC-Rio e Edições Loyola, 2011.

PRINCE, Susan. *Antisthenes of Athens*: texts, translations, and commentary. Ann Arbor: University of Michigan Press, 2015.

PROCLO. *Lecturas del Crátilo de Platón*; edición de Jesús M. Álvarez Hoz, Ángel Gabilondo Pujol y José M. García Ruiz. Madrid: Akal, 1999.

XENOFONTE. Memoráveis. Tradução do grego, introdução e notas de Ana Elias Pinheiro. Coimbra: Imprensa da Universidade de Coimbra, 2009.

Recebido em: 05-03-2019

Aceito para publicação em: 30-07-19

EMOÇÕES CORPORIFICADAS: UMA PERSPECTIVA SISTÊMICA SOBRE ESTADOS EMOCIONAIS[1]

EMBODIED EMOTIONS: A SYSTEMIC PERSPECTIVE ON EMOTIONAL STATES

MATHEUS DE MESQUITA SILVEIRA[2]
Universidade de Caxias do Sul (UCS) – Brasil
mdm.silveira@gmail.com

RESUMO: A tese principal deste artigo é que emoções são parcialmente constituídas por estados corporais. No centro da discussão está a tradição James-Lange e serão estabelecidas duas premissas para identificar as abordagens mais relevantes acerca desse ponto: (i) a Premissa de James, que será investigada a partir do trabalho do próprio autor; e (ii) a Premissa de Lange, que será investigada a partir do trabalho de Prinz. Embora se inscreva na mesma tradição, a perspectiva apresentada neste artigo diverge de ambas posições ao defender uma visão sistêmica, na qual as emoções são constituídas tanto por estados periféricos corporificados quanto por estados cerebrais. A visão sistêmica parece vulnerável a duas críticas principais: (i) de que as emoções, ao contrário de estados corporais, possuem intencionalidade; e (ii) as neurociências mapearam as bases neurais das emoções, então não há motivo para colocar fora do cérebro os componentes que as constituem. Em relação à primeira crítica, argumenta-se que as emoções não possuem intencionalidade direta, uma vez que não influenciam os comportamentos de modo específico, como objetos ou conteúdos representacionais o fazem. No tocante à segunda crítica, as conclusões das neurociências dizem menos do que aparentam acerca dos constituintes das emoções. A confusão na literatura empírica resulta da falha em distinguir corretamente emoções de desejos. Isso será realizado mediante a análise do debate entre as teorias de James-Lange para, posteriormente, mostrar como a Premissa de Lange desconstrói teorias relevantes desta posição, inclusive a sua própria.

PALAVRAS-CHAVE: Emoções. Estados corporais. Estados cerebrais. James. Lange.

ABSTRACT: *The main thesis of this article is that emotions are partly constituted by bodily states. The James-Lange tradition is key to this discussion, by considering emotions as bodily perceptions. In order to identify the most relevant general approaches to this subject, two premises will be established: (i) James' premise, which will be investigated from the author's work; and (ii) Lange's premise, which will be investigated from Prinz's work. Although the perspective presented in this article subscribes to the previously mentioned tradition, it diverges from both premises by the defense of a systemic view in which emotions are constituted by both embodied peripheral states and brain states. The systemic view seems vulnerable to two main critiques: (i) that emotions, unlike bodily states, have intentionality, and (ii) that neuroscience has mapped the neural basis of emotions, so there is no reason to place their constituent components outside the brain. Regarding the first criticism, it is argued that emotions do not have direct intentionality, since they do not influence the behaviors in the particular way that both objects and representational contents do. As for the second critique, neuroscience's conclusions show less than they seem to about the constituents of the emotions. The confusion in the empirical literature results from failure to correctly distinguish emotions from desires. This will be accomplished by analyzing the debate between James-Lange's theories, to present how Lange's premise undermines relevant theories of those positions, including his own.*

KEYWORDS: *Emotions. Bodily states. Brain states. James. Lange.*

[1] O presente artigo foi desenvolvido com apoio da FAPERGS/CAPES, a partir do edital 03/2018.
[2] Programa de Pós-Graduação em Filosofia da Universidade de Caxias do Sul (UCS).

Introdução

O ponto central de discussão deste artigo se refere à possibilidade de estados emocionais serem reduzidos a percepções corporais. James (1922) foi um dos primeiros pesquisadores a sistematicamente considerar essa hipótese, defendendo que percepções desse tipo compõem a natureza única das emoções. Nesta investigação, essa posição será identificada como a *Premissa de James*. Caso essa hipótese seja aceita, então é preciso considerar que os estados emocionais consistem nas expressões associadas a uma ou mais emoções, ou seja, em percepções corporais. Essa posição será chamada de *visão perceptiva*. Todavia, caso essa hipótese seja negada, então é preciso investigar a possibilidade de compreender os estados emocionais por meio de outras posições teóricas.

Uma posição semelhante à *Premissa de James* foi defendida por Lange (1922). O autor defende que alterações no corpo influenciam as emoções, evidenciando-se pelo uso de substâncias psicoativas ou redução da ansiedade após um exercício matinal. Para o autor, a melhor explicação para tais efeitos é a de que as emoções são redutíveis a estados periféricos corporificados[3]. Nesse sentido, o argumento de que a alteração desses estados pode desencadear emoções será denominado *Premissa de Lange* e essa posição será chamada de *visão periférica*. A questão é que, caso os estados emocionais sejam mediados por percepções corporais, então essa posição poderá ser interpretada como afirmativa no que toca ao papel da causalidade dentro de uma fenomenologia das emoções.

A *Premissa de Lange* se contrapõe às posições que identificam intrinsecamente emoções com atitudes proposicionais, como a apresentada por muitas teorias cognitivistas e da apreciação[4]. A dificuldade dessas posições está em explicar por que variações específicas nos estados periféricos desencadeiam determinadas alterações emocionais. A vantagem da *Premissa de James* é que ela apresenta uma explicação consistente com a perspectiva naturalista. O autor afirma que uma fenomenologia das emoções está circunscrita por percepções corporais, as quais produzem as características fisiológicas inerentes aos estados emocionais. Entretanto, essa posição diverge da *Premissa de Lange* no que concerne à questão de as emoções terem como efeito as percepções corporais ou, pelo menos, serem idênticas a elas. Portanto, ambas as premissas apontam na direção de que apenas uma está correta; a questão é como resolver o problema. Desse modo, um ponto importante discutido neste artigo é o de que a *Premissa de Lange* desconstrói teorias das emoções relevantes dentro desta tradição, incluindo a sua própria.

Neste artigo, será proposta uma nova perspectiva sobre as emoções, que parte da *visão periférica* e visa responder aos problemas identificados nesta posição teórica. O caminho parte de uma revisão de literatura da psicologia

[3] Considerando que o cérebro é uma estrutura corporal, na discussão acerca da corporificação para além do cérebro será utilizado o termo *estado periférico*, no lugar de *corpo* ou *estado corporal*. Da mesma forma, será utilizado o termo *percepção corporal* para se referir às percepções dos próprios estados periféricos do sujeito.

[4] Embora não seja objeto de discussão deste artigo, o leitor que desejar maior aprofundamento neste ponto poderá verificar que Solomon (1988) e Nussbaum (2001) são comumente colocados como cognitivistas, enquanto Ellsworth e Scherer (2003) estão identificados com uma teoria da apreciação.

cognitiva social e das neurociências que embasam a *Premissa de Lange*, de modo a usar as mesmas bases empíricas das críticas mais proeminentes à tradição James-Lange. Isso será realizado a partir da discussão da teoria desenvolvida por Prinz (2004), antes de recorrer ao próprio trabalho de Lange (1922). Na parte final deste artigo será apresentada uma variação da *visão periférica*, que receberá o nome de *visão sistêmica*, argumentando-se que é a teoria mais adequada para explicar a *Premissa de Lange* do que outras dessa tradição.

1 *Premissa de Lange* e a literatura empírica das emoções

As evidências apresentadas por Lange (1922), que defende a possibilidade de alterar ou incitar emoções mediante a manipulação de estados periféricos, estão embasadas majoritariamente em observações pessoais. Contudo, há um campo diversificado de pesquisas empíricas que reforçam sua posição. Nesta seção serão examinadas evidências dessa natureza, de modo que seja possível discutir posteriormente suas implicações às teorias das emoções apresentadas neste artigo.

O primeiro ponto desta parte será demonstrar que alterações no estado periférico incitam emoções. A análise das expressões faciais tem sido a metodologia mais utilizada em pesquisas sobre manipulação periférica. Essa perspectiva foi desenvolvida inicialmente por Darwin (1872), ao sustentar que as alterações fisiológicas causadas pelas emoções geram um impacto direto nelas, ao invés de simplesmente serem uma consequência. Em linhas gerais, a ideia é que os movimentos faciais podem influenciar os estados emocionais.

> A expressão livre mediante sinais externos de uma emoção a intensifica. Por outro lado, a repressão, tanto quanto possível, de todos os sinais externos suaviza nossas emoções [...] Até mesmo a simulação de uma emoção tende a despertá-la em nossas mentes (DARWIN, 1872, p. 366)[5].

Strack et al. (1988) realizaram um experimento, no qual manipularam as expressões faciais dos participantes enquanto eles ouviam histórias, ao solicitar que segurassem um lápis na boca, de modo a forçar a expressão de um sorriso ou uma carranca. O resultado foi que, sem saber que manifestavam tais expressões, julgaram como mais divertidas as histórias quando o lápis os forçava a sorrir. Com a intenção de replicar os resultados do experimento acima, Soussignan (2002, p. 70) coloca que "um efeito congruente entre o Sorriso de Duchenne[6] e os eventos que provocam emoções foi claramente apoiado pela autorreferência da experiência emocional e, parcialmente, pela reatividade do sistema nervoso autônomo". Em outras palavras, descobriu-se que apenas esse sorriso específico desencadeou uma emoção.

Evidências empíricas também são encontradas em pesquisas com diferentes abordagens sobre os estados periféricos. Stepper e Strack (1993) demonstraram

[5] Todas as traduções no decorrer do texto são do autor.
[6] O *Sorriso de Duchenne* envolve a contração do músculo zigomático principal e do músculo orbicular dos olhos. Esse sorriso pode ser metaforicamente descrito como um sorrir com os olhos.

que a postura corporal exerce influência na manifestação do orgulho, enquanto Philippot et al. (2002) identificaram padrões de respiração distintos relacionados a alegria, raiva, medo e tristeza. Os pesquisadores demonstraram que a indução de processos respiratórios é suficiente para incitar as respostas emocionais correspondentes. Esse é o mais intuitivo dos resultados apresentados até o momento, uma vez que diversas áreas da psicologia recomendam com solidez a utilização de padrões respiratórios para lidar com estados emocionais específicos.

As neurociências também fornecem evidências acerca dos estados periféricos. Adolphs et al. (2000) relatam que danos no córtex somatossensorial, diretamente associado à percepção corporal, prejudicam o reconhecimento visual de diversas expressões emocionais. Segundo Berntson et al. (2011), lesões na ínsula podem acarretar uma menor excitação e reação emocional como resposta a diferentes estímulos. Nessa linha, Craig (2009) coloca que a ínsula é ativada durante experiências emocionais e está diretamente conectada com os sistemas cerebrais associados à motivação.

> A amígdala pode não ser necessária para determinar se e em que medida um estímulo é desejável ou aversivo, hostil ou acolhedor; em vez disso, pode ser importante para registrar a excitação ou o impacto emocional de um estímulo (e especialmente dos estímulos aversivos). Em contraste, o córtex insular parece estar mais amplamente envolvido no reconhecimento, processamento e atribuição de valência avaliativa, e pode contribuir à excitação afetiva (BERNTSON et al., 2011, p. 85).

Tomadas sob uma perspectiva ampla, essas evidências suportam duas possíveis conclusões. A primeira conclusão é que as análises acerca da influência das expressões faciais e posturas corporais nas emoções, combinadas com as descobertas sobre o papel da ínsula e do córtex insular nos estados emocionais, apontam que os estados periféricos produzem efeitos motivacionais inerentemente emocionais. A segunda conclusão consiste no fato de que tanto as pesquisas sobre as funções da ínsula e do córtex insular quanto as do córtex somatossensorial conferem plausibilidade à posição de que os estados periféricos desempenham um papel central na autoconsciência das próprias emoções.

A literatura empírica não apenas reforça a *Premissa de Lange* como abre espaço para que uma perspectiva corporificada seja central a essa posição. De fato, a *visão periférica* permite que se considere que as emoções consistam em alterações padronizadas no corpo, porém enraizadas em sistemas cerebrais evolutivamente ancestrais e compartilhados por inúmeras espécies. Sendo assim, evidências empíricas que atestem uma relação entre alterações fisiológicas e emoções não apenas contribuirão na defesa de uma perspectiva evolucionista à questão como também reforçarão o argumento naturalista sobre o papel das emoções enquanto motivadoras de comportamento. A posição de que estados periféricos produzem efeitos motivacionais com base nas emoções está direcionada à confirmação da *Premissa de Lange*. Já a perspectiva de que os sistemas cerebrais possibilitam que os estados periféricos sejam centrais à autoconsciência emocional

consiste na pedra de toque à avaliação da especificidade periférica. Esse segundo ponto será o foco de investigação da próxima seção deste artigo.

2 Emoções e especificidades periféricas

Um problema empírico importante, enfrentado tanto por teorias da percepção corporal como dos estados periféricos, pode ser denominado de especificidade periférica. A questão colocada é que, caso emoções sejam corporificadas, então na *visão periférica* cada uma estaria correlacionada com um tipo diferente de estado periférico. A *visão perceptiva* também requer essa correlação pois, caso contrário, as percepções dos estados periféricos não seriam distintas entre si. Em ambas as visões, é necessário que ocorra uma variação suficiente entre os estados periféricos relevantes para que seja possível explicar variações entre as emoções.

O primeiro passo para responder esse problema é esclarecer como as emoções podem ser individualizadas. A *visão periférica* afirma que toda emoção é um tipo específico de estado periférico. No entanto, a pergunta é se esse posicionamento realmente está comprometido com a visão de que todo estado periférico consiste em um tipo de emoção. Essa relação é necessária, pois de outro modo não seria possível distinguir emoções de outros tipos de estados periféricos. O ponto é que a *visão periférica* precisa considerar todas as emoções associadas às percepções corporais, pois, caso não o faça, seria impossível distingui-las de outras formas de percepções desta natureza.

Acredita-se que a melhor forma de abordar essa questão é a partir da funcionalidade das emoções. As abordagens dentro da linha teórica estabelecida por James-Lange que não relacionam as emoções com estados cognitivos representativos de relações com o meio externo como, por exemplo, o perigo ou a perda, foram definidas por Smith e Lazarus (1990) como ancoradas em *temas relacionais centrais*. Todavia, as teorias corporificadas poderiam sustentar que a funcionalidade das emoções consiste no auxílio para lidar efetivamente com as circunstâncias externas ao sujeito. Portanto, elas seriam despertadas exatamente pelos estados corporais e módulos cerebrais cuja função é detectar situações deste tipo.

> A tarefa da apreciação para a pessoa é avaliar as circunstâncias percebidas em termos de um número relativamente pequeno de categorias de significância adaptativa, correspondendo a diferentes tipos de benefício ou dano, cada um com diferentes implicações para o enfrentamento. Uma proposição chave de uma teoria cognitivo-relacional da emoção é que o processo de avaliação resulta na identificação de uma relação de pessoa-ambiente molar, ou o que chamamos de "tema relacional nuclear", e que cada tema distinto resulta numa emoção distinta (SMITH e LAZARUS, 1990, p. 617).

De modo a realizar uma possível distinção entre emoções e estados periféricos, é necessário apenas determinar se um estado periférico específico é

despertado como resposta a uma circunstância particular. Somente em caso afirmativo é que poderia ser caracterizado como emocional. O problema é que tanto as percepções corporais quanto os estados periféricos como, por exemplo, a sede, não podem ser simplesmente definidos como emoções. No caso citado, sua funcionalidade está associada ao equilíbrio interno do organismo mediante a hidratação, e não à prestação de auxílio relativo a alguma condição externa.

A partir do exposto acima, considera-se necessário que teorias corporificadas relacionem determinadas emoções com manifestações específicas de estados periféricos. O problema está caso existam dois estados periféricos para uma mesma emoção e vice-versa. Entretanto, sustenta-se que uma incompatibilidade não é motivo suficiente para abandonar toda a abordagem, uma vez que o ponto problemático pode simplesmente estar na forma pela qual se está individualizando as emoções. Por exemplo, caso o ciúme e a raiva possuam os mesmos estados periféricos, pode-se propor que não exista uma diferença intrínseca entre ambas as emoções. Da mesma forma, o medo pode estar relacionado a dois estados periféricos diferentes, sendo plausível a proposta de que ele simplesmente seja manifestado por mais de uma expressão. Para realizar essas classificações, seria necessário um método mais preciso sobre como individuar as emoções e os estados periféricos. A questão é que as abordagens advindas exclusivamente da literatura conceitual são contraditórias e não apresentam uma solução clara para esse ponto.

Considerando as incongruências que permeiam a discussão estritamente conceitual deste problema, dados empíricos tornam-se interessantes guias de investigação. Os estados periféricos mencionados pelas teorias corporificadas das emoções são controlados pelo sistema nervoso periférico, que é dividido entre os sistemas nervosos autônomo e somático. Esse último é responsável pelos músculos passíveis de serem controlados voluntariamente, como os associados a expressões faciais, vocalizações, níveis de tensão muscular e movimentos corporais coordenados, ou seja, todos os que apresentam variação na manifestação de emoções. Por sua vez, o sistema nervoso autônomo controla a forma e a intensidade da excitação corporal.

Kreibig (2010, p. 394) relata uma "considerável especificidade de resposta do sistema nervoso autônomo na emoção ao considerar subtipos de emoções distintas". De fato, a autora salienta que um estado de excitação emocional aumentará em correlação com a frequência cardíaca, a transpiração e o fluxo sanguíneo em diferentes partes do corpo. Segundo Folkow (2000), as alterações simpáticas tendem a aumentar o grau de excitabilidade, enquanto as parassimpáticas o diminuem. Em outras palavras, mudanças fisiológicas viscerais estão associadas com emoções.

> [...] cada órgão e tecido é inervado por vias simpáticas e parassimpáticas distintas, sugerindo que cada órgão e tecido tem alguma independência funcional [...] Considera-se que esta independência pertence ao sistema medular simpático-adrenal, para o qual há uma influência direta do sistema nervoso, e uma influência de hormônios adrenomedulares que, na maioria das

> situações, têm diferentes papeis funcionais (STORBECK e WYLIE, 2018, p. 202).

Argumenta-se que não há necessidade das teorias dentro da linha de James-Lange demonstrarem a existência de uma especificidade autonômica ou somática. Considerando como é formado o sistema nervoso periférico, de fato há menor gasto energético se duas ou mais emoções estiverem correlacionadas ao mesmo estado autonômico, desde que também o estejam com relação a diferentes estados somáticos e vice-versa. Contudo, pesquisas sobre a especificidade periférica têm metodologicamente se concentrado em analisar ambos os estados separadamente, motivo pelo qual o mesmo será realizado neste artigo.

Em defesa da especificidade somática, Ekman e Friesen (1971) argumentam que existem expressões faciais universais intrínsecas às emoções básicas. Em sua pesquisa, os autores apresentam que mesmo os participantes que não possuíam contato com a cultura ocidental associavam expressões faciais específicas as mesmas emoções. Dessa perspectiva decorre a defesa de que as emoções básicas de medo, raiva, tristeza, nojo, surpresa e alegria estão associadas a expressões universalmente correspondentes e independem de fatores socioculturais.

> A experiência dentro de uma cultura, os tipos de eventos que tipicamente provocam emoções particulares, podem influenciar a capacidade de discriminar pares específicos de emoções. Os rostos de medo podem não terem sido distinguidos dos rostos de surpresa, porque nessa cultura os eventos de medo também são quase sempre surpreendentes; isto é, o súbito aparecimento de um membro hostil de outra aldeia, o encontro inesperado de um fantasma ou feiticeiro, etc (EKMAN e FRIESEN, 1971, p. 129).

A crítica que pode ser feita é que os resultados apresentados acima se devem a uma escolha de paradigma forçada, na qual é conferido aos participantes um conjunto limitado de emoções para escolher. Russell (1994, p. 116) coloca que "não foi dada aos participantes a opção de dizer nenhuma das alternativas acima [...] Eles não tinham autorização para descrever o rosto com parte de uma resposta instrumental [...]". O autor apresenta que os resultados encontrados mediante metodologias de livre escolha são mais variados. No entanto, mesmo em métodos de escolha forçada, é relevante o dado associado à invariabilidade cultural entre expressões faciais e emoções.

Expressões faciais podem estar inter-relacionadas com alterações somáticas e autonômicas, o que possibilitaria a distinção coletiva de diferentes emoções. Prinz (2004) enfatiza que percepções corporais não necessitam que mudanças nos estados somáticos estejam perfeitamente correlacionadas com emoções específicas. Em geral, as teorias corporificadas não sustentam a necessidade de emoções serem exclusivamente inatas. Entende-se que o seu fundamento biológico não é reducionista, uma vez que a cultura pode influenciar parcialmente as emoções, que são biologicamente aptas a receber estes estímulos.

> É possível que as características que compõem as expressões faciais sejam desencadeadas por sistemas de resposta inatos, enquanto

> todo o conjunto depende do fato de que as condições que provocam estas respostas muitas vezes ocorrem conjuntamente (PRINZ, 2004, p. 88).

Pesquisas sobre alterações de estados autonômicos associados à frequência cardíaca, à pressão arterial e à resposta galvânica de pele também são relevantes em pesquisas sobre emoções. Ekman et al. (1983) demonstram que determinadas mudanças autonômicas são correlatas às emoções específicas. Todavia, Cacioppo et al. (2000) ressaltam que essa correlação é distinguível apenas no tocante a grupos de emoções. A raiva, o medo e a tristeza apresentam maior alteração na frequência cardíaca e menor atividade vascular que o nojo, da mesma forma que este apresenta maior resposta galvânica de pele que a alegria. Ainda que os resultados empíricos não sejam uniformes, Sequeira et al. (2009) salientam que grande parte dos dados reforçam a hipótese de que emoções possuem assinaturas autonômicas distintas.

A crítica a uma abordagem corporificada das emoções pode ser realizada salientando o vasto número de pesquisas com resultados inconclusivos ou disformes entre si. Por sua vez, a defesa pode ressaltar que há dificuldades metodológicas advindas de limitações tecnológicas para realizar uma quantificação empírica e, mesmo assim, há dados que apontam a existência de correlações entre emoções e estados periféricos. O desenvolvimento de melhores métodos de individuação e maior precisão indutiva acerca das emoções certamente trarão uma nova luz à questão. Entretanto, neste momento não é possível tirar conclusões empíricas fortes acerca da especificidade periférica. Dessa forma, será defendida neste artigo uma abordagem corporificada com base em outros alicerces, que serão apresentados na próxima seção.

2.1 Emoções e estímulos periféricos

Pesquisas sobre os efeitos de lesões na medula espinhal sobre emoções conferem uma nuance interessante à discussão sobre emoções corporificadas. Cannon (1927) defende que pacientes nessas condições não apresentam alterações nas suas capacidades de sentir emoções. Como resposta, abre-se a possibilidade de as emoções serem apenas percepções corporais. Hohmann (1966, p. 149) defende que a emoção "é um tipo de pensamento mental em vez de uma sensação dirigida fisicamente". Contudo, o próprio autor ressalta que lesões medulares mitigam as emoções, ainda que não as eliminem.

> Elas podem ser controladas, no entanto, pelos processos no córtex cerebral, por processos condicionados por todos os tipos de impressões anteriores. O córtex também pode controlar toda a maquinaria periférica, exceto as vísceras. Os processos inibitórios no tálamo não podem colocar o organismo em ação, exceto as partes que não estão sob controle voluntário, mas a agitação lá pode produzir emoções da maneira usual, e possivelmente com maior violência por causa da inibição (CANNON, 1927, p. 123).

Chwalisz et al. (1988) analisaram pacientes com lesão na coluna vertebral e relataram que eles apresentaram emoções intensas. No que toca ao medo, os autores demonstram que a pontuação dos pesquisados com relação ao grupo de controle foi maior. Essa divergência pode ocorrer caso os participantes do experimento de Hohmann (1966) tenham oferecido relatos imprecisos acerca do que estavam sentindo. Em uma perspectiva histórica, práticas terapêuticas encorajaram a adoção de posturas estoicas com relação a lesões desse tipo, o que pode resultar em um efeito de contraste nas pesquisas. Afinal, o trauma da lesão pode ser tão intenso que outras emoções são mitigadas por comparação. As próprias condições de sobrevida do período impediam esses pacientes de possuir qualidade de vida, levando-os a um ostracismo emocional. A mudança nessa perspectiva pode explicar o porquê de pacientes nessas condições terem apresentado maior intensidade em suas experiências emocionais. O ponto é que nenhuma das pesquisas apresentadas refutam categoricamente perspectivas somáticas das emoções.

Não há concordância dentro da literatura empírica sobre essa questão, o que torna relevante averiguar se há pontos de convergência nas pesquisas sobre o tema. A pesquisa de Heims et al. (2004) teve como foco pacientes com falha autonômica em vez de lesão na medula espinhal. Os autores descobriram que eles participavam normalmente do *Iowa Gambling Task*[7], mas falhavam ao atribuir emoções a personagens de histórias curtas. Não houve alteração nas respostas somáticas dos participantes, ainda que apresentassem deficiências autonômicas. Na verdade, isso explica o porquê de suas reações não serem particularmente intensas. A questão é que a evidência empírica de que uma falha autonômica dificulta a atribuição de emoções oferece suporte à *visão perceptiva* e à *visão periférica*. No escopo das teorias tradicionais sobre as emoções, torna-se difícil compreender o motivo de uma deficiência autonômica interferir em uma emulação empática efetiva.

> Não ocorreu interação do grupo com a emoção [...]. No entanto, a ANOVA revelou um efeito maior no grupo [...] com pacientes de performance pior que o grupo de comparação. O desempenho fraco dos pacientes nessa tarefa sugere que repostas autônomas corporais podem ter um papel na previsão, talvez por emulação empática, dos estados de sensibilidade emocional subjetivos de outros (HEIMS et al., 2004, p. 1984).

O problema consiste em explicar o motivo pelo qual emoções não são extinguidas por lesões da medula espinhal. Quanto a isso, Prinz (2004) atenta para

[7] O *Iowa Gambling Task* é um teste psicológico pensado para simular tomadas de decisões da vida real. Foi introduzido inicialmente por Bechara et al. (1994) e ganhou notoriedade quando Damásio (1994) propôs a hipótese dos marcadores somáticos. De fato, o autor defende que as emoções usam o corpo como seu teatro e afetam o modo como vários circuitos cerebrais operam. Nesse sentido, a variedade das respostas emocionais exerce um papel causal nas mudanças tanto nos estados corporais quanto nos cerebrais. Para Damásio (1999), o conjunto destas mudanças subsidiam os padrões neurais que se tornam a sensação consciente da emoção. Embora não seja o escopo deste artigo, com relação ao caso apresentado no texto, sugere-se que o desempenho dos sujeitos não se deve a questões emocionais, mas a desejos deficientes. Nesse sentido, uma deficiência emocional não influenciaria, necessariamente, o desempenho no teste.

a possibilidade de haver estímulos periféricos no rosto ou em outras partes do corpo não afetados por lesões desse tipo. Essa posição torna possível a uma abordagem corporificada explicar a experiência emocional relatada por pacientes com essa debilidade, pois a diferença é que seria estimulada por um escopo mais estreito dos estados periféricos.

> Evidências da indução corporal da emoção, lesão da medula espinhal e neuroimagem sugerem que emoções são a consequência causal de mudanças corporais. Elas são estados que registram mudanças corporais. Caso este seja o caso, então mudanças corporais devem ser capazes de causar emoções. Isto não significa que toda emoção é o resultado de alguma perturbação anterior no corpo, mas sugere que as perturbações são causas confiáveis das emoções (PRINZ, 2004, p. 58).

Outra possibilidade é que as emoções poderiam resultar de uma alucinação relativa aos estados periféricos. Na *visão perceptiva*, isso significa a possibilidade de emoções reais resultarem de percepções alucinatórias. A *visão periférica*, ao contrário, teria que defender a inexistência de qualquer emoção real, pois do delírio dos estados periféricos decorreria necessariamente a alucinação emocional.

A literatura empírica sobre a relação entre estímulos periféricos e lesões corporais está longe de ser conclusiva. A posição que encontra maior convergência é a de que há alguma interferência no funcionamento das emoções, sempre que os estados periféricos forem incapazes de enviar estímulos ao cérebro. Até o momento, a perspectiva na qual toda emoção depende desses estímulos não é corroborada nem refutada. Uma vez que as evidências empíricas são incapazes de falsear as perspectivas corporificadas, essa posição teórica é reforçada. Na pendência de evidências ou melhores argumentos de refutação, a questão a ser abordada passa a ser o modo como as emoções podem ser influenciadas por estímulos periféricos.

3 Emoções e conteúdo representacional

No escopo da abordagem de James-Lange, grande parte das teorias estão próximas da *visão perceptiva*. Isso ocorre por elas considerarem que alterações periféricas resultam de diferentes emoções. A questão levantada nesta seção é se elas são capazes de explicar como e por que ocorrem essas mudanças. Dada sua correlação e aprimoramento da *visão perceptiva*, a teoria de Prinz (2004) será tomada como representativa da *Premissa de James*.

Contemporaneamente, a *visão perceptiva* ganhou contornos psicossemânticos. Prinz (2004, p. 190) coloca que "emoções envolvem estados corporais, mas não representam estados corporais. Elas usam estados corporais para representar as relações organismo-ambiente". O autor defende que um módulo de calibração consiste em um grupo de sistemas cerebrais cuja função é causar as alterações periféricas idênticas às relacionadas com determinada emoção. O medo de baratas é um bom exemplo para explicar esse ponto. A percepção do inseto ou de seus rastros faz com que se conclua que ele está próximo. A

percepção da barata seria uma crença de existência cujo conteúdo é **há uma barata por perto**. Toda vez que o inseto estiver presente, ou a sua lembrança, a percepção correspondente registrada no módulo de calibração desencadeará um estado periférico associado ao medo.

A reação apresentada acima não corresponde ao medo, mas a uma percepção associada a um estado periférico. A *visão perceptiva* possui um aspecto contraintuitivo, uma vez que o medo tem na barata a sua referência, e não a aceleração da frequência cardíaca ou qualquer outra alteração periférica. O problema é como equalizar o medo como uma percepção corporal, e a resposta está em uma teoria causal do conteúdo representacional.

> Nesta abordagem, um estado carrega informação sobre aquilo com o que ele ocorre simultaneamente de forma confiável. Nos casos mais típicos, a ocorrência simultânea em questão é causal. Um estado carrega informações sobre o que faz com que ele ocorra de maneira confiável. A causação confiável não implica correlação perfeita. Dizer que uma coisa é causada de forma confiável por outra coisa significa apenas que, caso aquela segunda coisa ocorresse, então, todas as coisas sendo iguais, a primeira teria uma alta probabilidade de ocorrer como resultado (PRINZ, 2004, p. 53).

O conteúdo representacional de um objeto surge mediante a apreensão de informações sobre suas características. Sempre que o conteúdo apreendido representa propriedades do objeto, então elas podem ser consideradas como o conteúdo nominal da representação. Diferentemente, o conteúdo real consiste nas características essenciais da classificação natural dentro da qual determinado objeto se encontra. No momento em que o conteúdo representacional é composto de informações apreendidas diretamente do conteúdo real, então essas características consistem no conteúdo real da representação.

O conteúdo nominal de uma representação corresponde a alterações pontuais nos estados corporais. Por exemplo, o medo decorre da percepção de se encontrar em um estado periférico específico. O conteúdo real desta percepção pode ser expresso como **estou em perigo**. A literatura dentro da *visão perceptiva* coloca essa relação como um tema relacional nuclear. Lazarus (1991, p. 121) define que "um tema relacional nuclear é simplesmente o dano ou benefício central (consequentemente nuclear) em encontros adaptativos que subjazem cada tipo específico de emoção". Há inúmeras relações benéficas e cada uma delas constitui um tema relacional nuclear que desencadeia uma emoção positiva distinta, como a alegria associada à conquista de um objetivo. A abordagem defendida por Prinz (2004) consiste exatamente na defesa de que o conteúdo das emoções são os temas relacionais centrais.

A ideia é que o conteúdo representacional possa rastrear os conteúdos reais tendo como ponte conteúdos nominais, mas isso levanta um problema acerca das emoções. No momento em que elas não detectam necessariamente um tema relacional nuclear mediante a sua descrição, então também não poderiam detectar sua essência. Todavia, disso não decorre que as emoções estejam ausentes de estrutura ou que consistam em meros indicadores. A resposta a essa questão passa

por uma aproximação entre uma teoria da apreciação e a *visão perceptiva*. Em linhas gerais, as emoções rastreariam temas relacionais nucleares a partir do registro de alterações nos estados corporais.

A *visão perceptiva* coloca como possível que qualquer propriedade essencial de um animal possa ser classificada como o conteúdo representacional associado a ele. Prinz (2002, p. 149) exemplifica que se "alguém está procurando por cães na tundra ártica, alguém pode invocar a representação de um típico cão de trenó". Nesse caso, as características desses animais consistem apenas no conteúdo nominal do conceito humano associado a **cães de trenó**. Para o autor, a referência ao conteúdo real ocorre a partir do registro dos conteúdos representacionais e nominais, com os sistemas cerebrais presentes nos módulos de calibração separando os dois últimos. Prinz (2004, p. 68) destaca que "os detectores de rastreamento de essência têm conteúdo real e conteúdo nominal, mas com os detectores de rastreamento de aparência, estes se separam". Dessa forma, o conteúdo representacional de **cães de trenó** está vinculado aos estímulos periféricos reais causados por esses animais.

O argumento tem como premissa que emoções são percepções corporais, o que parece contraditório com a alegação de que elas consistem em temas relacionais nucleares. A tentativa de distensionar ambas as posições se dá pela defesa de que emoções representam temas relacionais nucleares a partir da percepção de alterações corporais. Por exemplo, o medo rastreia o perigo através de palpitações cardíacas da mesma forma que o conceito de **cão de trenó** o faz por meio das características físicas desses animais. Dessa forma, os temas relacionais nucleares são o conteúdo real das emoções, enquanto as alterações corporais consistem no seu conteúdo nominal.

> A evolução, sem dúvida, nos dotou com respostas fisiológicas distintas para várias situações que nossos ancestrais encontraram. O coração está predisposto a acelerar (junto com diversas outras respostas fisiológicas) quando percebemos objetos iminentes, cobras, insetos rastejantes, ou sombras grandes se movendo na noite; ou quando ouvimos sons altos ou gritos específicos; ou quando sentimos o cheiro de um predador. O coração acelerado e outras alterações fisiológicas que ocorrem coletivamente nestas condições servem como um detector de perigo (PRINZ, 2004, p. 69).

A correlação entre temas relacionais nucleares e percepções corporais é possível devido ao modo como fora estabelecida sua conexão. Percepções corporais de perigo evoluíram de modo a engatilhar as mudanças fisiológicas apropriadas, sendo que esta conexão pode ser tanto genotípica quanto fenotípica. Independentemente de sua matriz, repostas ao perigo estão baseadas em um sistema filogenético primitivo que predispõe determinadas reações com relação a eventos dessa natureza.

O conteúdo das emoções consiste em temas relacionais nucleares, cabendo ao estado periférico informar a sua existência. Nos módulos de calibração estão contidas informações sobre temas relacionais nucleares específicos, que causam a

ocorrência de percepções corporais através de sistemas cerebrais. Emoções consistem nessas percepções corporificadas na medida em que elas representam determinado tema relacional nuclear mediante a modificação de um estado periférico. Nesse sentido, uma emoção é uma reação visceral. Considerando que temas relacionais nucleares aparentemente seriam de importância vital para a sobrevivência, a funcionalidade das emoções seria tornar consciente a sua existência.

3.1 Emoções e percepções corporais

A relação entre a *visão perceptiva* e os temas relacionais nucleares pode ser criticada por descaracterizar o papel das percepções corporais. Uma forma de compreender esse ponto é a partir do exemplo de alguém que sente um aperto no peito e fraqueza nas pernas ao saber que perdeu o emprego. Essas reações não necessitam da consciência acerca das dificuldades futuras para serem sentidas. A percepção do estado periférico é posterior às alterações corporais, pois ela simplesmente torna consciente o estado de ansiedade ou medo existente. Essa crítica é plausível, uma vez que a própria tentativa de evitar a expressão dessas emoções requer esforço e possui um alto custo energético. Ao contrário da teoria apresentada na seção anterior, no exemplo acima as percepções corporais não tornam conscientes os temas relacionais nucleares, mas as próprias emoções.

No entanto, a capacidade de percepções corporais tornarem conscientes temas relacionais nucleares não pode ser descartada apenas pelo argumento do parágrafo anterior. Essas associações estão presentes na cultura popular, como no caso de um arrepio na nuca significar uma situação de perigo. Ainda que tenham algum grau de confiabilidade, é preciso uma relação mais forte do que esses relatos para inferir um tema relacional nuclear com base em uma determinada emoção. Acerca das emoções, Prinz (2004, p. 66) diz que "[...] elas são causadas de forma confiável por propriedades relacionais que pertencem ao bem-estar. Mas a representação requer mais do que a causação confiável". Fisiologicamente, a piloereção é um correlato periférico do medo e os sistemas cerebrais que provocam essa reação têm a função de tornar consciente uma situação perigosa. A crítica é que se está inferindo o acesso a temas relacionais nucleares baseados em emoções a partir de crenças relacionais, o que não confere qualquer objetividade natural à relação.

Uma possível resposta a essa crítica é que ela não faz qualquer menção a uma abordagem psicossemântica. Afinal, sempre que a piloereção traz à consciência o perigo, a percepção corporal também o faz. Entretanto, tornar consciente determinada situação é insuficiente e Prinz (2004) é claro na defesa de que conteúdos representacionais estão vinculados a apreensão de informações. Uma vez que essa funcionalidade deriva de uma história evolutiva, torna-se menos plausível que temas relacionais nucleares sejam acessados diretamente.

> É suficiente dizer que uma representação mental é um estado mental que é confiavelmente causado por algo e foi estabelecido evolutivamente ou por aprendizado para detectar esta coisa.

> Colocando mais concisamente, uma representação mental é um estado mental que foi *criado* para ser *desencadeado* por algo (PRINZ, 2004, p. 54).

O argumento é o de que conteúdos representacionais são estabelecidos de diferentes formas. Eles podem resultar do aprendizado, como no conceito de cão de trenó; ou do processo evolutivo, como em sistemas cujas células respondem à presença de perigo. Dretske (1988, p. 107) argumenta que eles "exibem as propriedades essenciais de crenças genuínas: eles têm um conteúdo proposicional, e sua posse desse conteúdo auxilia a explicar por que o sistema em que eles ocorrem se comporta da maneira como o faz". Nesse caso, os conteúdos representacionais poderiam ser engatilhados por situações diferentes das associadas à sua função original. Por exemplo, a capacidade de visualizar rachaduras finas em uma escada resulta de uma adaptação evolutiva relacionada a evitar quedas em precipícios. Burge (1986, p. 26) defende que "estas representações (percepções) são formadas por processos que são relativamente imunes à correção de outras fontes de informação; e as representações da visão inicial parecem ser totalmente independentes da linguagem". O módulo mental para detectar defeitos em superfícies permanece inalterado, mas sua gama de aplicação pode ser expandida.

> É possível que os modelos representacionais se apliquem de maneira diferente, sem que as reações físicas ou os poderes discriminatórios sejam diferentes. Estes fatos, junto com o fato de que muitos eventos e estados mentais fundamentais são individualizados em termos dos modelos representacionais relevantes, são suficientes para gerar a conclusão de que muitos eventos e estados mentais não são individualisticamente individualizados: eles podem variar enquanto o corpo e os poderes discriminatórios da pessoa são concebidos como constantes (BURGE, 1986, p. 27).

A crítica ao argumento apresentado acima é que ele não oferece uma fundamentação sólida que fortaleça a hipótese de que representações corporais têm como função representar temas relacionais nucleares. Millikan (1993, p.11) coloca que "quando funciona adequadamente, uma representação mental ocorre simultaneamente com o que está representado, retrata o que se representa, e... participa em inferências apropriadas". Os módulos cerebrais desempenham esse papel em condições normais e estão ancorados em sistemas do cérebro herdados evolutivamente. Nessa perspectiva, a funcionalidade das percepções corporais é apenas trazer à consciência emoções que surgem como resposta a contextos situacionais contemporâneos. Em uma perspectiva psicossemântica, essa relação se encontra em uma corporificação das emoções, e não em temas relacionais nucleares.

> Retratar, indicar e inferir estão igualmente envolvidos na representação humana, mas como normas biológicas e não como meras disposições. Não são os fatos sobre como o sistema opera que o tornam um sistema representativo e determinam o que ele

> representa. Ao contrário, são os fatos sobre o que estaria fazendo se estivesse operando de acordo com as normas biológicas (MILLIKAN, 1993, p. 10).

Há um ponto de discórdia, pois enquanto algumas posições consideram um erro a ativação de módulos cerebrais por contextos situacionais diferentes dos originais, outras colocam que isso nada mais é do que um desvio de função que aumenta o seu valor adaptativo. Essa discussão levanta o problema acerca do porquê percepções corporais representam emoções. A primeira alternativa de resposta considera que os estados periféricos podem compor as emoções de modo parcial ou integral. Dessa forma, uma percepção corporal é equivalente a uma emoção. A segunda possibilidade de resposta coloca que estados periféricos são causados por emoções e, consequentemente, são distintos delas. Essa posição é compatível com as teorias cognitivistas convencionais, pois a manifestação de estados periféricos é uma característica das emoções.

3.2 Emoções e fatores de valência

A segunda crítica à *visão perceptiva* defende que reforços internos são qualificadores de valência emocional. Abordagens comportamentais tendem a colocar que reforços são estímulos externos, definindo-os com relação à probabilidade de motivarem comportamentos futuros. Prinz (2004) defende esses reforços como internos, porém sem abandonar o viés comportamental, concordando que eles são definidos mediante seu impacto em ações vindouras. Essa perspectiva coloca que uma valência negativa está associada a um reforço interno negativo, da mesma forma que uma valência positiva está vinculada a um reforço interno positivo. Assim, a valência emocional pode ser vista como a expressão de maior ou menor negatividade ou positividade.

> Identificar a valência com reforços internos explica por que as emoções às vezes impulsionam a aproximação ou o recuo. As emoções negativas nos encorajam a recuar das situações que as provocam, e emoções positivas nos encorajam a buscar as situações que as provocam (PRINZ, 2004, p. 174).

A distinção entre desejo e valência é que o primeiro representa um estado emocional associado com atos de fala imperativos, enquanto a segunda é um imperativo da ação. Por exemplo, o medo possui uma valência negativa, não apenas tornando consciente uma situação perigosa, mas motivando um comportamento de fuga. Prinz (2004, p. 174) sustenta que "[...] os RPIs [reforços positivos internos] e os RNIs [reforços negativos internos] servem como imperativos internos. Um RPI serve como um comando que diz algo como 'mais disso!', enquanto o RNI diz 'menos disso!'". O autor defende que uma valência positiva ou negativa comporta a presença de um desejo da mesma polaridade. Em suma, a *visão perceptiva* coloca as emoções em duas polaridades; a percepção positiva de algo e o desejo que persista, e a percepção negativa de algo e o desejo que termine.

A *visão perceptiva* sustenta que as emoções são percepções corporais dotadas de um fator de valência e, por isso, podem ser interpretadas como uma teoria sobre a valência das percepções corporais. Prinz (2004, p. 164) defende que a "[...] valência é uma característica real da nossa psicologia e é essencial à emotividade". Dessa forma, torna-se necessário que a valência emocional das percepções corporais motive a continuidade ou o encerramento de algo a partir de sua polaridade. A questão é que isso ocorre mesmo em casos não associados diretamente com emoções. A sensação da pressão dos dedos durante uma massagem possui valência positiva e estimula o desejo de que continue. A valência de percepções corporais tem a mesma funcionalidade para além das emoções, sendo que em nenhum caso perde o fator motivacional.

O problema consiste no fato de que a valência emocional deveria influenciar a resposta ao tema relacional nuclear. Smith e Kirby (2009, p. 104) reforçam essa relação ao sustentarem que "[...] cada emoção distinta tem seu próprio tema relacional nuclear distinto, o qual representa um tipo particular de relacionamento adaptativo às circunstâncias de alguém". Pode-se argumentar que as percepções corporais motivam a continuidade e a cessação de emoções e temas relacionais nucleares. A acrofobia é um exemplo, pois o sofrimento durante um episódio de medo intenso contém um fator motivacional suficiente para que se evite escalar montanhas. A valência negativa do medo está motivando a prevenção do perigo sob condições específicas, ou seja, evita-se o perigo para não sentir medo. A questão é que essa posição é mais consistente com a defesa de que a valência emocional estimula diretamente o desejo para prolongar ou encerrar emoções ao invés dos temas relacionais nucleares.

> O resultado é que as emoções parecem conter duas partes distintas. Toda emoção tem uma valência, a qual compartilha com outras emoções. Porém, toda emoção também tem um perfil corporal distinto refletido por diferenças (às vezes sutis) na atividade neuronal (PRINZ, 2004, p. 163).

Caso a valência emocional não corresponda às percepções corporais, isso enfraquece a *visão perceptiva*. O problema é que, para que a valência emocional estimule o desejo pela continuidade ou encerramento de temas relacionais nucleares, é necessário que ela seja diferente das percepções corporais, pois estas o fazem com relação apenas às emoções. A crítica é que o argumento central da *Premissa de James* falha em sua tentativa de abarcar nuances importantes dessa questão.

Há grande plausibilidade na colocação de que o medo motiva o afastamento do perigo, enquanto a raiva leva ao seu confronto. Lerner e Keltner (2001) sustentam que a raiva motiva comportamentos de risco, da mesma forma que o medo eleva o grau de reticência diante de ações dessa natureza. O argumento é que essas emoções influenciam o comportamento em um escopo maior do que apenas o dos temas relacionais nucleares. Schnall et al. (2008) sustentam essa posição, defendendo que o nojo (emoção com valência negativa) confere maior severidade aos juízos morais.

> [...] os efeitos que encontramos do nojo nos juízos morais não são meramente a manifestação de uma tendência geral para que o afeto negativo amplifique os juízos morais. Assim, parece que quanto mais claramente os participantes estão sentindo nojo, mais diretamente esse afeto é tomado como atribuição para os juízos morais (SCHNALL et al. 2008, p. 12).

A relação do nojo com o aumento da severidade de juízos morais explica por que comportamentos socialmente danosos são apreendidos como fonte de contaminação social. Contudo, não é tão clara a relação de como, por exemplo, a tristeza abranda juízos dentro da esfera da moralidade. Emerge desse ponto a necessidade de um modelo sobre como emoções estimulam desejos e, consequentemente, motivam comportamentos, sem negar ou distorcer os diferentes dados dentro da literatura empírica.

> [...] toda emoção parece ter o que pode ser chamado de "marcador de valência". Neste modelo revisado, a tristeza é um estado composto, contendo tanto uma apreciação que detecta a perda, quanto uma ativação que representa a perda como algo negativo (PRINZ, 2004, p. 163).

Há uma considerável literatura empírica sobre os aspectos motivacionais das emoções. Dutton e Aron (1974, p. 516) concluem seu experimento defendendo "[...] a noção de que a emoção forte, em si, aumenta a atração sexual do sujeito para com sua contraparte feminina". A emoção identificada pelos autores foi o medo, que estimulou a atração sexual em condições situacionais ansiogênicas. Por sua vez, Isen e Levin (1972) reportam que a valência emocional positiva é motivacional com relação a comportamentos sociais.

> Os resultados dos dois estudos tomados em conjunto fornecem suporte à noção de que se sentir bem leva a ajudar. Porque o sentir-se bem tem sido gerado em uma variedade de formas e configurações, e desde que o tipo de medida de ajuda e a fonte das populações estudadas também variaram, essa relação parece ter alguma generalidade empírica (ISEN e LEVIN, 1972, p. 387).

O ponto é que essas pesquisas apenas arranham a superfície da questão, pois cada emoção apresenta como efeito seu próprio perfil motivacional. Por uma questão de clareza argumentativa, esses efeitos serão denominados impulsos genéricos. A escolha deriva do fato de que emoções distintas motivam comportamentos de diversas maneiras, enquanto aquelas pertencentes ao mesmo estado emocional parecem motivar ações de formas semelhantes. Entretanto, não há consenso empírico acerca do funcionamento dos impulsos genéricos. A convergência de resultados é um pouco maior com relação ao fato de que cada emoção fortalece ou enfraquece temporariamente os desejos. A partir desse ponto, pode-se estabelecer que a diferença entre o medo e a raiva é que enquanto o primeiro fortalece desejos direcionados a evitar confrontos, a segunda estimula comportamentos na direção oposta. Todavia, essa distinção abrange todo o escopo dessas emoções, pois a raiva fortalece o desejo de agir agressivamente e o medo

o faz com relação ao sexo. A hipótese mais plausível é de que alterações de estímulos nos desejos são proporcionais a duração e a intensidade da emoção.

A primeira questão acerca dos impulsos genéricos consiste em explicar como as emoções podem influenciar o comportamento de modos opostos. Por exemplo, o medo pode estimular o desejo de não ser percebido pelo componente dominante de um grupo ao mesmo tempo que pode motivar a aproximação em direção a ele. Frijda (1986, p. 83) argumenta que "a situação meramente provoca a ação; a prontidão da ação existe apenas na medida em que a inibição pode bloquear a execução da ação". À medida que as motivações da ação são fixas e rígidas, a ideia de tendência de comportamento perde muito do seu significado.

Esse problema não surge na concepção de impulsos genéricos. Na medida em que há flexibilidade no modo como emoções influenciam comportamentos, noções como tendência comportamental e prontidão à ação ganham maior peso e significado. No exemplo anterior, em ambos os casos o medo estimula o desejo de evitar o risco. Uma forma natural de evitar o perigo em um grupo em que o componente dominante é um aliado motiva a aproximação, mas quando há incerteza ou, ainda, a certeza de que ele representa um desafeto, não ser percebido é a melhor forma evitar qualquer dano. O ponto em comum entre esses comportamentos é o impulso genérico associado à aversão ao risco.

> Programas flexíveis são aqueles que são compostos por cursos alternativos de ação, que permitem a variação de circunstâncias e respostas nas ações executadas. Em tais programas, anseios, intenções e objetivos tornaram-se independentes das ações particulares. [...] Com tal estrutura, é significativo falar da emoção (FRIJDA, 1986, p. 83).

A segunda questão acerca dos impulsos genéricos está na possibilidade de resistir a eles. Da ideia de que um desejo é estimulado pela intensidade emocional não decorre que ele sempre conduzirá a um comportamento específico e descontrolado; o que é um traço de grande valor adaptativo. Por exemplo, a resistência ao nojo permite que se limpe um bebê, e o autocontrole da raiva pode evitar um confronto letal. O ponto é que resistir a um impulso genérico não significa que ele não seja sentido. O autocontrole em um episódio de raiva pode evitar o confronto físico, entretanto, pode se manifestar em tremores musculares, de modo que sua repressão também possa gerar fadiga física e emocional. O importante aqui é que a heteronomia das respostas comportamentais e a ausência ocasional de expressão das emoções são compatíveis com a ideia de impulso genérico.

A pergunta central aqui é se a *visão perceptiva* possui a mesma compatibilidade. Enquanto concebidas como percepções corporificadas, as emoções podem influenciar o comportamento de duas formas, pelo seu conteúdo ou sua valência. Prinz (2004, p. 174) salienta que "[...] marcadores de valência podem impactar tanto o comportamento presente quanto o futuro. Eles influenciam o futuro em virtude da sua habilidade de influenciar o presente". O ponto é que a valência emocional não estimula a continuidade ou encerramento

dos temas relacionais nucleares, conforme é defendido pelo autor. A crítica é que a *visão perceptiva* não consegue, amparada em fatores de valência, explicar o porquê da raiva (valência negativa) motivar tanto o encerramento de uma discussão ofensiva quanto a inclinação para entrar em um confronto; da mesma forma que o medo (valência negativa) motiva tanto a fuga do perigo quanto estimula a atração sexual.

Para avançar nessa questão, é insuficiente estabelecer que percepções corporais representam emoções em vez de temas relacionais nucleares. Em geral, percepções são acompanhadas de atitudes mentais assertivas, cuja função está em afirmar mentalmente algo. Uma vez que impulsos genéricos alteram a intensidade dos desejos, é necessário explicar por que afirmar mentalmente um estado emocional sistematicamente induz uma predisposição a determinados comportamentos.

3.3 Emoções e os estados periféricos

Após analisar os problemas enfrentados pela *visão perceptiva*, é intelectualmente honesto realizar o mesmo com a *visão periférica*. Embora a *Premissa de Lange* não seja alvo das mesmas objeções, o argumento é que ela é igualmente insuficiente para responder às questões levantadas neste artigo. Serão realizadas duas objeções principais a essa visão, com ambas as críticas direcionadas à defesa de que percepções corporais mediam a relação entre os estados periféricos e centrais.[8]

A *visão periférica* parte da premissa de que estados periféricos engatilham percepções corporais que produzem as características comportamentais e motivacionais qualitativas das emoções. James (1922, p. 65) defende que a emoção "não pode existir, contudo, sem seus atributos físicos". Em suma, a *Premissa de Lange* identifica as emoções com os estados periféricos, com sua força motivacional sendo mediada por percepções corporais. O argumento é que a manipulação direta da percepção corporal produz os mesmos impulsos genéricos suscitados pelos estados periféricos. Uma vez que ela produza efeitos emocionais específicos, parece inegável que estes sejam considerados como emoções genuínas.

> De fato, não é difícil provar agora, e por meio das experiências mais comuns e ordinárias, que as emoções podem ser induzidas por uma variedade de causas que são totalmente independentes das perturbações da mente, e que, por outro lado, elas podem ser suprimidas puramente por meios físicos (LANGE, 1922, p. 66).

A *Premissa de Lange* estabelece que, ao manipular as percepções corporais para induzir uma emoção que ocorre em certo contexto situacional, também se estará estimulando o estado periférico que constitui determinada emoção. Uma vez que esse estado periférico não seja real, a *visão periférica* caracteriza essa reação

[8] O objetivo desta classificação é destacar o contraste entre os estados cerebrais. Nesse sentido, estados centrais são governados pelo sistema nervoso central, enquanto os estados periféricos são regidos pelo sistema nervoso periférico.

como alucinatória. Lange (1922, p. 71) esclarece que "o frenesi causado por um cogumelo, por exemplo, ou por mania, pode ter a mesma aparência de raiva, mas não é a verdadeira raiva, tão pouco como a alegria que vem de beber vinho não é alegria real [...]". A crítica é que em todos os casos mencionados os impulsos genéricos são idênticos. Desse modo, os defensores da *visão periférica* precisam explicar por que a manipulação de percepções corporais não produzem emoções genuínas.

Conforme colocado no início desta seção, a segunda objeção também é baseada na concepção da *visão periférica* de que percepções corporais mediam os efeitos dos estados periféricos. Lange (1922, p. 80 – tradução do autor) ressalta que "se as impressões que recaem sobre os nossos sentidos não possuíssem o poder de estimulá-los, nós vagaríamos pela vida sem simpatia e paixão [...]". A crítica consiste no fato de que caso as percepções corporais engatilhem os impulsos genéricos mediante um erro na transmissão da informação, então os estados periféricos não poderiam desempenhar melhor essa função, pois seus efeitos são mediados por essas percepções. Nesse quesito, a crítica apresentada anteriormente à *visão perceptiva* também se aplica à *visão periférica*.

> Caso a concepção da natureza dos afetos aqui representados seja estabelecida, então podemos esperar que toda influência envolvendo mudanças gerais no sistema nervoso vascular deve ter uma expressão emocional. Evidentemente, não podemos esperar que essas emoções coincidam exatamente com o fenômeno para o qual, geralmente, nós reservamos essa denotação; as diferenças na causa resultarão naturalmente em vários efeitos (LANGE, 1922, p. 68 – tradução do autor).

Na seção anterior, argumentou-se que a valência emocional de percepções corporais não produz impulsos genéricos. No momento em que a *visão periférica* coloca a percepção corporal como mediadora da emoção, a valência apenas influencia o desejo de modo que esta seja prolongada ou encerrada. O argumento é que a consciência da emoção possibilita a preferência pela sua permanência ou não, sendo que isso a diferencia de estados alucinógenos ou patológicos, pois apenas no primeiro caso ela denota corretamente a relação com o fenômeno real. A crítica é que essa distinção não explica por que emoções produzem impulsos genéricos.

Na *Premissa de Lange*, caso os estados periféricos produzam impulsos genéricos, estes seriam mediados por estados centrais a partir do conteúdo informacional recebido mediante percepções corporais. Lange (1922, p. 79) é categórico ao afirmar que "as emoções não são forças que estão fora do corpo [...]". Nesse caso, o problema está na possibilidade de os estados periféricos não influenciarem os desejos. Na *visão periférica*, as melhores candidatas para desempenhar essa função são as percepções corporais. No entanto, uma vez que elas não produzem impulsos genéricos, então não há um mediador adequado aos estados periféricos e, consequentemente, eles também não podem desempenhar essa função.

Em linhas gerais, há duas formas de resolver essa questão. Na primeira possibilidade é possível argumentar que o estímulo periférico é mediado por outro tipo de estado que, oposto à percepção corporal, é adequado para produzir impulsos genéricos. Esse argumento necessita de três etapas centrais. A primeira consiste na distinção entre percepções corporais e estados centrais não perceptuais, que receberiam os estímulos periféricos. A segunda se refere a uma possível existência de estados centrais não perceptuais, que receberiam os estímulos dos estados periféricos relevantes. A terceira está direcionada a necessidade de esses estados serem mais adequados do que as percepções corporais para produzir os impulsos genéricos.

As duas primeiras etapas não são controversas, enquanto a terceira apresenta o problema central. Na possibilidade apresentada, os estados centrais só seriam diferentes das percepções corporais se representassem estados periféricos. O ponto é que, considerando que eles recebem estímulos exatamente desses estados, não fica claro qual seria sua outra função. É possível argumentar que os estados periféricos possuem o perfil psicossemântico que a *visão perceptiva* atribui às percepções corporais, de modo que eles representassem temas relacionais nucleares. Todavia, isso levanta novos problemas. O primeiro é que não há evidências empíricas que corroborem essa hipótese. Já o segundo, conforme fora argumentado anteriormente, consiste no fato de que um estado qualquer que represente um tema relacional nuclear não é adequado para produzir impulsos genéricos. Em uma perspectiva evolutiva, pode-se acrescentar ainda que isso não proporciona qualquer vantagem adaptativa ou economia energética.

A segunda possibilidade para resolver essa questão seria argumentar que as percepções corporais são facilitadoras dos impulsos genéricos. Além de ser contrário à primeira possibilidade, esse argumento é triangular. Nesse caso, a defesa da *visão periférica* necessita que percepções corporais possam facilitar a produção de impulsos genéricos. O ponto é que essa possibilidade enfraquece a primeira crítica contra a *visão perceptiva* desenvolvida anteriormente. A menos que as percepções corporais possam ser facilitadoras dos impulsos genéricos, tanto a *visão perceptiva* quanto a *visão periférica* possuem problemas. Por outro lado, caso fosse possível (o que não tem nenhum respaldo empírico), então os argumentos das seções anteriores favoreceriam a *Premissa de James* com relação à *Premissa de Lange* que, de qualquer forma, fica enfraquecida.

4 Estados emocionais e a *visão sistêmica*

A *visão periférica* fica enfraquecida quando confrontada com abordagens corporificadas mais atuais das emoções. O argumento principal apresentado na parte final deste artigo é que uma *visão sistêmica* sobre as emoções não possui as mesmas vulnerabilidades. O ponto de partida é o mesmo da *Premissa de Lange*, que coloca os estados periféricos como parte das emoções, com a diferença de que também se inclui os estados centrais como necessários à sua existência. Considera-se essa abordagem sistêmica precisamente por colocar os sistemas nervosos central e periférico como fundamentais aos estados emocionais.

Estabelecer a relação entre os sistemas nervosos central e periférico, assim como suas respectivas funcionalidades, é fundamental para esclarecer a *visão sistêmica*. Os componentes periféricos têm duas funções principais: preparar o corpo para realizar determinadas ações e sinalizar estados emocionais. Os componentes centrais têm a função primária de produzir impulsos genéricos e demais efeitos psicológicos. Ordinariamente, ambos atuam em unidade para manter a funcionalidade do organismo. Por exemplo, quando o sistema nervoso periférico reage agressivamente e sinaliza pela expressão corporal a prontidão para ações de confronto, essa informação é transmitida ao sistema nervoso central, que produz um impulso genérico que estimula esse comportamento. Da mesma forma, quando o sistema nervoso central produz um impulso genérico que estimula a agressividade, essa informação é transmitida ao sistema nervoso periférico, que prepara o corpo para a ação, sinalizando-o pela expressão corporal.

Mesmo entre posições cognitivistas, não é controverso que estados periféricos sejam os responsáveis pela expressão corporificada de emoções. A controvérsia está na premissa de que alterações periféricas sejam estímulos motivacionais, uma vez que essa função seria exclusiva do sistema nervoso central. O problema ocorre quando é colocado que um estímulo motivacional necessita de uma reação corporal correspondente, pois as emoções são essenciais na comunicação interpessoal e isso requer a existência de sinalizadores externos. Griffiths e Scarantino (2009, p. 5) colocam que "emoções têm sido vistas como respostas mais ou menos precisas sobre como as coisas são, mas elas também são e, talvez principalmente, mais ou menos respostas eficazes orientadas para objetivos". A flexibilidade oferecida por essa abordagem coloca no nível biológico a contribuição das emoções com relação à aptidão social e à funcionalidade do grupo, enquanto elementos externos, como a cultura, estabelecem gatilhos que não existiam no contexto natural. Nesse caso, a relação é possível apenas porque existe uma estrutura biológica apta a receber as influências do ambiente, sendo flexível o suficiente para moldar e ser moldada por ele.

> Em contraste, uma perspectiva situada sobre a emoção enfatiza o papel do contexto social na produção e no gerenciamento de uma emoção, e a influência recíproca da emoção no contexto social em evolução. Comportamentos que tradicionalmente têm sido vistos como expressões involuntárias do estado psicológico do organismo são vistos como sinais projetados para influenciar o comportamento de outros organismos ou como "movimentos" estratégicos em uma transação contínua entre organismos (GRIFFITHS e SCARANTINO, 2009, p. 4).

O componente social das emoções tem um aspecto estratégico negligenciado nas abordagens cognitivistas dentro do escopo teórico desenvolvido por James-Lange. Já a *visão sistêmica* defende essa funcionalidade mediante uma relação entre os sistemas nervosos central e periférico, que é estabelecida por reforços mútuos. Morris e Keltner (2000, p. 20) reforçam que "[...] emoções podem ser vistas como tendo funções - não porque foram projetadas, mas porque foram selecionadas com base em sua adaptabilidade [...]". Defende-se aqui a utilidade de expressar intenções quando se está genuinamente motivado a realizar determinada

ação e há condições físicas para isso. O perigo está em expressar a intencionalidade sem que exista motivação ou capacidade para se comportar adequadamente. Por esse motivo é fundamental que exista sincronia entre a motivação e a modulação das reações corporais.

> Empiricamente, a sugestão de dividirmos a operação do programa dos afetos da "emoção cognitiva superior" parece ignorar [...] a extensão em que a tomada de decisões de "ordem superior" tem que se valer do sistema límbico para funcionar (BLACKBURN, 1998, p. 129).

A exigência de uma integração sistêmica efetiva para o funcionamento de organismos complexos não é algo estranho à literatura empírica em biologia. Afinal, todos os mamíferos superiores possuem essas características bem documentadas. A questão não é que exista relação entre os sistemas nervosos central e periférico, mas que estes estejam identificados com as emoções. Justifica-se a defesa da *visão sistêmica* pela sua capacidade de resolver de maneira mais simples os problemas levantados nas discussões atuais acerca tanto da *visão perceptiva* quanto da *visão periférica*.

4.1 Componentes centrais da *visão sistêmica*

O ponto principal desta seção é identificar os componentes centrais da *visão sistêmica* a partir da sua capacidade de produzir impulsos genéricos. Sustenta-se que esses componentes estão associados a um módulo cerebral central vinculado ao sistema nervoso central, responsável pela produção de impulsos dessa natureza. O objetivo central será definir os módulos cerebrais envolvidos nesse processo de forma mais precisa. O caminho a ser trilhado é o de estabelecer maior comprometimento com as questões balizadoras do debate contemporâneo sobre os estados emocionais.

Os componentes centrais da *visão sistêmica* são identificados como estados emocionais. Eles produzem estímulos motivacionais inerentemente emocionais, com a particularidade de fazê-los de forma difusa. Esse ponto é problemático em relação às teorias que colocam a intencionalidade a um objeto como necessária aos estados emocionais. Crane (1998, p. 234) defende que “as dores normalmente parecem ter localização e extensão no espaço e tempo, e nós falamos delas sem esforço usando termos singulares e predicamos propriedades delas como fazemos a objetos e eventos”. A ideia é que as emoções se direcionam a objetos específicos, ainda que não se tenha consciência deles. Solomon (1976) apresenta uma resposta, defendendo que estados emocionais estão identificados com objetos de alcance geral.

> Uma emoção não é distinta ou separável de seu objeto; o objeto como um objeto desta emoção não tem existência separado da emoção. Há dois componentes, minha raiva e o objeto da minha raiva [...] toda emoção tem a forma unitária de "minha-emoção sobre...", "estou com raiva de..." [...] A emoção é distinguida pelo seu objeto; não há nada além do seu objeto. Mas também não há tal objeto sem a emoção (SOLOMON, 1976, p. 178).

Essa abordagem coloca os componentes centrais dos estados emocionais como uma classe especial de desejo ou aversão, pois seriam constituídos por atitudes mentais do mesmo tipo. A emoção do medo passa a ser uma aversão a um mal generalizado, da mesma forma que a tristeza seria com relação à perda. Por exemplo, a visualização de uma aranha engatilha uma aversão expressa pelo medo de ser atacado, produzindo também uma aversão generalizada ao risco de sofrer danos físicos e engatilhando um processo que resulta em um estado periférico específico. Da mesma forma, outras aversões ou desejos são capazes de engatilhar processos que produzirão novos estados periféricos que serão influenciados por inúmeros fatores externos.

O problema é que essa abordagem não explica a característica episódica das emoções e, consequentemente, os efeitos da manipulação periférica. É um ponto pacífico estabelecer que nojo e raiva podem coexistir, mas o mesmo não se aplica à alegria. O nojo ou a alegria podem sobrepujar um ao outro, mas nenhum será inteiramente dominante em estados emocionais compostos por emoções opostas. A coexistência harmônica entre a raiva e o nojo ocorre porque ambos estão no mesmo escopo de valência. Isso não se deve a intencionalidade ou relação com um objeto específico, mas simplesmente a constituição biológica de ambos. Dessa forma, no nojo provocado por manipulação periférica, o impulso genérico é compatível ao manifestado pela raiva, e ambos são opostos ao encontrado na alegria. Essa oposição ocorre porque o cérebro de mamíferos superiores não comporta permanentemente a simultaneidade dos estados emocionais associados a emoções de valência oposta, desencadeando um colapso nervoso caso uma não sobrepuje a outra.

> É verdade que a maioria das pessoas não se torna emocional para cumprir alguma obrigação social. Mas, uma análise dos papéis não é mais objetável a este respeito do que uma análise em termos de padrões adaptativos baseados na biologia [...] qualquer episódio específico de raiva, amor, ... etc. pode não atender a nenhuma necessidade social. Mas, caso em média ou a longo prazo tais síndromes emocionais estiverem em conformidade com as normas sociais, então seu resultado líquido será funcional dentro do sistema social (AVERILL, 1980, p. 336).

A intencionalidade não é necessária para compreender a funcionalidade dos estados emocionais. Griffiths (1997, p. 148) coloca que "a resposta emocional é produzida 'estrategicamente' com a intenção de extrair uma resposta adequada dos outros. Mas a resposta emocional é interpretada pelo agente e sua comunidade como natural e involuntária". Nessa linha, Sizer (2000) também argumenta que os componentes centrais das emoções não possuem conteúdo intencional.

> Mudanças nas operações neste nível (funcional) têm efeitos globais nos estados e processos do nível de representação, mas de uma forma que é independente do conteúdo semântico destes estados. As mudanças são de baixo para cima e, portanto, influenciam a criação e o desempenho de todas as representações e processos representacionais afetados. Isto é o que explica os efeitos

> generalizados e independentes de conteúdo do humor (SIZER, 2000, p. 763).

A *visão sistêmica* relaciona não apenas os estados emocionais com aspectos funcionais, mas também o faz com relação a seus componentes centrais. As emoções engatilham impulsos genéricos e alteram temporariamente os parâmetros na ordem dos desejos e aversões. A questão vantajosa dessa abordagem é que ela explica com maior simplicidade a abrangência dos estados emocionais. Isso ocorre porque seus componentes centrais não constituem uma mera representação dentro de um sistema, mas um estado corporificado que modela as representações dentro dele. Dessa forma, é possível que emoções contrárias apresentem, enquanto componentes parciais dos estados emocionais, efeitos diacrônicos dentro de um mesmo organismo.

Conclusão

Em grande parte, a estrutura geral da visão sistêmica se apoia na unificação dos temas centrais à discussão com aqueles que se encontram relegados a um segundo plano. O presente artigo apresentou uma questão central à visão sistêmica e ao modo como emoções podem ser compreendidas, ou seja, que a manipulação periférica altera as emoções. Esse ponto não é especificamente controverso, mas é considerado por muitas abordagens como um fenômeno marginal. As mudanças periféricas alteram as motivações pelo mesmo motivo que as alterações motivacionais modificam os estados periféricos, pois os sistemas responsáveis pela motivação, preparação da ação e expressão estão integrados. Em suma, as emoções são os estados que resultam da integração entre esses sistemas.

Dado o exposto, conclui-se que há suporte na literatura empírica à ideia de que a manipulação periférica influencia as emoções, de modo que apresenta consequências emocionais genuínas. Grande parte dos pesquisadores interessados no estudo das emoções utiliza esse ponto para defender a tradição James-Lange contra diversas teorias contemporâneas sobre o tema. Buscou-se apresentar que esse fundamento também pode ser utilizado para criticar posições dentro desse escopo teórico. É fato que há um longo caminho para se desenvolver de maneira mais completa os componentes centrais que integram os estados emocionais. Em suma, o objetivo deste artigo foi realizar um estudo exploratório inicial acerca dessa possibilidade, de modo a indicar que não há impeditivos óbvios para o desenvolvimento integral de uma *visão sistêmica*.

Referências

ADOLPHS, Ralph et al. A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *The Journal of Neuroscience*, v. 20, n. 7, p. 2685-2690, 2000.

AVERILL, James R. A constructivist view of emotion. In: PLUTCHIK, Robert e KELLERMAN, Henry (eds.). *Theories of emotion*. New York, NY: Academic, 1980. p. 305-339.

BECHARA, Antoine et al. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, v. 50, p. 7-15, 1994.

BERNTSON, Gary G. et al. The insula and evaluative processes. *Psychological Sciences*, v. 22, n. 1, p. 80-86, 2011.

BLACKBURN, Simon. *Ruling Passions: A theory of practical reasoning.* New York, NY: Oxford University Press, 1998.

BURGE, Tyler. Individualism and psychology. *Philosophical Review*, v. 95, p. 3-45, 1986.

CACIOPPO, John T., et al. The psychophysiology of emotion. In: LEWIS, Michael e HAVILAND, Jeannette M. (ed.). *Handbook of Emotions*. New York: Guilford Press, 2000. p. 173-191.

CANNON, Walter B. The James-Lange theory of emotion: A critical examination. *American Journal of Psychology*, v. 39, p. 106-124, 1927.

CHWALISZ, Kathleen, DIENER, Ed, e GALLAGHER, Dennis. Autonomic arousal feedback and emotional experience: Evidence from the spinal cord injured. *Journal of Personality and Social Psychology*, v. 54, p. 820-828, 1988.

CRAIG, A. D. How do you feel — now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, v. 10, p. 59-70, 2009.

CRANE, Tim. Intentionality as the mark of the mental. In: O'HEAR, Anthony (ed.). *Contemporary Issues in the Philosophy of Mind*. Cambridge: Cambridge University Press, 1998. p. 229-251.

DAMÁSIO, Antonio R. *Descartes' Error: Emotion, reason, and the human brain*. New York: Grosset/Putnam, 1994.

_____. *The feeling of what happens: Body and emotion in the making of consciousness.* New York: Harcourt Brace & Company, 1999.

DARWIN, Charles R. *The expression of the emotions in man and animals*. London, RU: John Murray, 1872.

DRETSKE, Fred. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press, 1988.

DUTTON, Donald G. e ARON, Arthur. Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, v. 30, n. 4, p. 510-517, 1974.

EKMAN, Paul e FRIESEN, Wallace. V. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, v. 17, p. 124-129, 1971.

_____ et al. Autonomic nervous system activity distinguishes among emotions. *Science*, v. 221, p. 1208-1210, 1983.

ELLSWORTH, Phoebe C. e SCHERER, Klaus. R. Appraisal processes in emotion. In: DAVIDSON, Richard J., SCHERER, Klaus. R. e GOLDSMITH, H. Hill (ed.). *Handbook of Affective Sciences*. New York, NY: Oxford University Press, 2003. p. 572-595.

FOLKOW, B Björn. Perspectives on the integrative functions of the 'sympatho-adrenomedullary system'. *Autonomic Neuroscience: Basic and Clinical*, v. 83, p. 101-115, 2000.

FRIJDA, Nico. *The emotions.* New York: Cambridge University Press, 1986.

GRIFFITHS, Paul. E. *What emotions really are.* Chicago, IL: University of Chicago Press, 1997.

_____ e SCARANTINO, Andrea. Emotions in the wild: The situated perspective on emotion. In: ROBBINS, Philip e AYDEDE, Murat (eds.). *Cambridge handbook of situated cognition*. New York, NY: Cambridge University Press, 2009. p. 437-453.

HEIMS, H. C. et al. Social and motivation functioning is not critically dependent on feedback of autonomic responses: neuropsychological evidence from patients with pure autonomic failure. *Neuropsychologia*, v. 42, n. 14, p. 1979-1988, 2004.

HOHMANN, George. W. Some effects of spinal cord lesions on experienced emotional feelings. *Psychophysiology*, v. 3, p. 143-156, 1966.

ISEN, Alice M. e LEVIN, Paula. F. Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology*, v. 21, n. 3, p. 384-388, 1972.

JAMES, William. What is an emotion? In: DUNLAP, Knight (ed.). *The emotions*. Baltimore: Waverly Press, 1922. p. 11-30.

KREIBIG, Sylvia D. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, v. 84, n. 3, p. 394-422, 2010.

LANGE, Carl G. The emotions. In: DUNLAP, Knight (ed.). *The emotions*. Baltimore: Waverly Press, 1922. p. 33-90.

LAZARUS, Richard S. *Emotion and adaptation.* New York, NY: Oxford University Press, 1991.

LERNER, Jennifer S. e KELTNER, Dacher. Fear, anger, and risk. *Journal of Personality and Social Psychology*, v. 81, p. 146-159, 2001.

MILLIKAN, Ruth G. *White Queen psychology and other essays for Alice.* Cambridge: MIT Press, 1993.

MORRIS, Michael W. e KELTNER, Dacher. How emotions work: The social functions of emotional expression in negotiations. *Research in Organizational Behavior*, v. 22, p. 1-50, 2000.

NUSSBAUM, Martha C. *Upheavals of thought: The intelligence of the emotions*, Cambridge, RU: Cambridge University Press, 2001.

PHILIPPOT, Pierre, CHAPELLE, Gaëtane, e BLAIRY, Sylvie. Respiratory feedback in the generation of emotion. *Cognition & Emotion*, v. 16, n. 5, p. 605-627, 2002.

PRINZ, Jesse J. *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, Massachusetts: MIT Press, 2002.

_____. *Gut reactions.* New York, NY: Oxford University Press. 2004.

RUSSELL, James. A. Is there universal recognition of emotion from facial expression? A review of cross-cultural studies. *Psychological Bulletin*, v. 115, p. 102-141, 1994.

SCHNALL, Simone et al. Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, v. 34, n. 8, p. 1096-1109, 2008.

SEQUEIRA, Henrique et al. Electrical autonomic correlates of emotion. *International Journal of Psychophysiology*, v. 71, n. 1, p. 50-56, 2009.

SIZER, Laura. Towards a computational theory of mood. *British Journal of the Philosophy of Science*, v. 51, p. 743-769, 2000.

SMITH, Craig A. e KIRBY, Leslie. D. Core relational themes. In: SANDERS, David e SCHERER, Klaus R. (eds.). *The Oxford Companion to Emotion and the Affective Sciences*. New York, NY: Oxford University Press, 2009. p. 104-105.

_____ e LAZARUS, Richard. S. Emotion and adaptation. In: PERVIN, Lawrence A. (ed.). *Handbook of personality: Theory and research*. New York, NY: The Guilford Press, 1990. p. 609-637.

SOLOMON, Robert C. *The passions.* New York, NY: Doubleday. 1976.

_____. On emotions as judgments. *American Philosophical Quarterly*, v. 25, n. 2, p. 183-191. 1988.

SOUSSIGNAN, Robert. Duchenne smile, emotional experience, and autonomic reactivity: A test of the facial feedback hypothesis. *Emotion*, v. 2, n. 1, p. 52-74, 2002.

STEPPER, Sabine e STRACK, Fritz. Proprioceptive determinants of affective and nonaffective feelings. *Journal of Personality and Social Psychology*, v. 64, n. 2, p. 211-220, 1993.

STORBECK Justin e WYLIE Jordan. The functional and dysfunctional aspects of happiness: Cognitive, physiological, behavioral, and health considerations. In: LENCH, Heather (ed.). *The Function of Emotions*. Springer, 2018. p. 195-220.

STRACK, Fritz, MARTIN, Leonard L. e STEPPER, Sabine. Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, v. 54, n. 5, p. 768-777, 1988.

Recebido em: 06-03-2019

Aceito para publicação em: 02-07-19