UNIVERSIDADE FEDERAL DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

PAULA FERNANDA PATRÍCIO DE AMORIM

A CRÍTICA DE JOHN SEARLE À INTELIGÊNCIA ARTIFICIAL:

uma abordagem em filosofia da mente

João Pessoa, PB

2014

PAULA FERNANDA PATRÍCIO DE AMORIM

A CRÍTICA DE JOHN SEARLE À INTELIGÊNCIA ARTIFICIAL:

uma abordagem em filosofia da mente

Dissertação apresentada ao Programa de Pós Graduação em Filosofia da Universidade Federal da Paraíba, para obtenção do título de Mestre em Filosofia.

Orientador: Prof. Dr. Ricardo Sousa Silvestre

João Pessoa, PB

2014

A524c Amorim, Paula Fernanda Patrício de.

A crítica de John Searle à inteligência artificial: uma abordagem em filosofia da mente / Paula Fernanda Patrício de Amorim.-- João Pessoa, 2014.

97f.

Orientador: Ricardo Sousa Silvestre

Dissertação (Mestrado) - UFPB/CCHL

1.Searle, John Rogers, 1932- crítica e interpretação. 2.Filosofia - crítica e interpretação. 3.Inteligência artificial. 4.Consciência. 5.Argumento do quarto chinês.

PAULA FERNANDA PATRÍCIO DE AMORIM

A CRÍTICA DE JOHN SEARLE À INTELIGÊNCIA ARTIFICIAL:
uma abordagem em filosofia da mente

Dissertação apresentada ao Programa de Pós Graduação em Filosofia da Universidade Federal da Paraíba, para obtenção do título de Mestre em Filosofia.

Aprovada em 27/02/2014

BANCA EXAMINADORA

_____

Prof. Dr. Ricardo Sousa Silvestre
Universidade Federal da Paraíba

_____

Prof. Dr. Giovanni da Silva de Queiroz
Universidade Federal da Paraíba

_____

Prof. Dr. Daniel Durante Pereira Alves
Universidade Federal do Rio Grande do Norte

Dedico este trabalho a todos aqueles que acreditaram no que eu acreditei, independentemente das difíceis condições de satisfação destas minhas crenças.

Os dualistas perguntaram: “Quantos tipos de coisas e propriedades existem?”, e contaram dois. Os monistas, confrontando-se com a mesma questão, chegaram somente até um. Mas o erro verdadeiro foi realmente começar a contar.

John Searle

AGRADECIMENTOS

Agradeço a todos os professores do Programa de Pós Graduação em Filosofia da Universidade Federal da Paraíba.

Ao coordenador do programa, Prof. Dr. Anderson D'Arc Ferreira, que pacientemente sempre deu suporte não só a mim, mas também a muitos alunos deste programa, quando mais precisamos.

Aos funcionários da coordenação do PPGFil, que muitas vezes sequer imaginam o quanto são importantes para o crescimento de um discente, prontamente auxiliando e buscando soluções dentro de turbilhões burocráticos. A eles, que nos ajudam, que nos ouvem, que nos aconselham e que tantas vezes excedem as obrigações de suas funções, fazendo muito mais por todos nós – tanto discentes quanto docentes.

Agradeço, sem que seja preciso citar nomes, a todos aqueles que me ajudaram de forma direta e indireta em prol da construção desse trabalho: das experientes indicações bibliográficas ao ombro amigo em momentos de desespero. A todos que se fizeram presentes, meu "muito obrigada".

RESUMO

Este texto é destinado a apresentar a crítica do filósofo John Searle à Inteligência Artificial, mais especificamente ao que Searle chama Inteligência Artificial Forte. Para isso, o principal texto pesquisado e esmiuçado será o seu artigo de 1980, *Minds, brains and programs,* no qual é apresentado o seu *argumento do quarto chinês.* O argumento busca demonstrar que não é possível duplicar a mente através de processos meramente formais, a saber, da manipulação de números 0 e 1 em um programa de computador, que poderia ser executado em qualquer computador com um *hardware* capacitado para executar esse tipo de programa. Para tanto, é sugerida por Searle uma experiência de pensamento que envolve o que ele chama de “quarto chinês”, que tem como objetivo demonstrar que nem um computador executando um programa que envolve a habilidade da compreensão jamais poderá compreender coisa alguma, já que em sua experiência de pensamento, os mesmos processos computacionais seriam emulados de em um ambiente diferente (o quarto, que seria o equivalente ao computador), por um ser humano (que seria o equivalente ao programa no computador) e ainda assim não seria possível afirmar ter havido compreensão alguma (nesse caso, do idioma chinês), nem por parte do quarto e nem por parte do homem, que estaria dentro do quarto realizando os mesmos processos que um computador rodando um programa realizaria, de acordo com Searle. Através da exposição do argumento mencionado e das críticas de Searle à Inteligência Artificial Forte, dentre outras teorias da mente, tais quais o computacionalismo e o funcionalismo, buscar-se-á atingir a compreensão do que seria a contribuição de Searle para a Filosofia da Mente, no que diz respeito às discussões em torno da Inteligência Artificial, analisando tanto o argumento quanto as suas críticas e os pontos fracos e fortes da argumentação searleana.

Palavras-chaves: inteligência artificial, consciência, argumento do quarto chinês.

ABSTRACT

This text is intended to present a critique of the philosopher John Searle on Artificial Intelligence, more specifically to what Searle calls Strong Artificial Intelligence. For this, the main researched and scrutinized text will be your 1980's article, Minds, brains and programs, in which is presented his chinese room argument. The argument seeks to show that it is not possible to duplicate the mind by purely formal processes, namely the manipulation of numbers 0 and 1 in a computer program that could be executed on any computer with a capable hardware to run this type of program. Therefore, it is suggested by Searle a thought experiment involving what he calls the "chinese room", which aims to demonstrate that even a computer running a program that involves the ability of understanding can never understand anything, since in his thought experiment, the same computational processes would be emulated in a different environment (the room, which will be equivalent to the computer), by a human being (which would be equivalent to the program on the computer) and still would not have been possible to state some understanding (in this case, the Chinese language), nor by the room neither by the man who would be within the room performing the same processes as a computer running a program accomplish, according Searle. Through exposure of that argument and criticism of Searle to Strong Artificial Intelligence, among other theories of the mind, such as the computationalism and functionalism, we will seek to achieve the understanding of what would be the contribution of Searle for the Philosophy of Mind with regard to discussions on Artificial Intelligence, analyzing both the argument as its criticism and the strengths and weaknesses of Searle's argumentation.

Key-words: artificial intelligence; consciousness; chinese room argument.

SUMÁRIO

INTRODUÇÃO

Em 1950, o matemático e lógico, Alan Turing publicou o seu texto intitulado *Computadores e inteligência*[1], contendo em seu corpo o tão conhecido *Teste de Turing*, que inicialmente foi nomeado pelo autor de "jogo de imitação". Além de um dos primeiros textos referentes à Inteligência Artificial, o artigo mencionado lança uma importante questão que até hoje intriga os filósofos da mente – seja para inspirar críticas ou para reformular conceitos: "as máquinas podem pensar?". Através de seu teste, Turing buscaria responder essa questão de forma afirmativa, defendendo que a inteligência não precisa ser uma característica unicamente humana.

Juntamente a essa questão, o que mais intrigou vários filósofos foi a problemática pressuposição de que um tipo de inteligência poderia ser desenvolvida a partir de algoritmos executados em artefatos eletrônicos criados pelo homem. Dentre estes, o filósofo americano John Searle se destaca a partir da sua tão repercutida publicação, em 1980, do artigo *Minds, brains and programs*[2]. O que tornou o seu artigo tão famoso foi o que ficou conhecido como *o argumento do quarto chinês*, o principal artifício utilizado por Searle em seu texto para elucidar a impossibilidade de duplicação da inteligência humana a partir de algorítmos.

Em seu texto, Searle elabora um exemplo buscando imitar os processos que ocorrem em uma máquina (o computador) a partir do momento em que ela está sendo programada para desempenhar uma dada ação e também enquanto esta ação está sendo realizada. Buscando externalizar os processos formais que são realizados no interior da máquina, o autor pretende demonstrar que ainda que esta execute bem o que lhe é pedido, recebendo informações e enviando respostas, ela continuaria incapacitada de *compreender* as informações com as quais estava lidando e as quais deveria transformar. Posto isto, o computador melhor programado poderia aparentar ter inteligência, mas na verdade não passaria de um artefato que obedece ordens (segue algoritmos) muito bem. O exemplo usado por Searle teria o propósito não apenas de contestar as posições de Turing, mas também de criticar teorias da mente que defendessem a possibilidade de um modelo computacional da mente, a partir de tão somente símbolos formais – em outras palavras, Searle

[1] TURING, Allan M. *Computing machinery and intelligence*. Mind, 1950.
[2] SEARLE, John R. Minds, brains and programs. In: _____. *Behavorial and brain sciences*. 1980.

buscava atingir o *computacionalismo* tanto quanto os posicionamentos de Turing. Em termos gerais, o maior alvo do argumento do quarto chinês é a Inteligência Artificial, ou o que Searle chama de Inteligência Artificial Forte, que seria um versão da Inteligência Artificial que acredita que é possível não apenas simular a mente humana em um ambiente artificial, a saber, em um computador digital; mais do que isso, essa versão defende a proposta de que é possível *duplicar* a mente humana. Searle defenderá que essa duplicação é impossível por diversos motivos, dentre eles o aspecto subjetivo do mental, como, por exemplo, a intencionalidade presente na mente humana e de alguns outros animais, sendo essa a base de sua crítica.

A crítica de Searle foi alvo de discussão em diversos centros acadêmicos (Berkeley, Yale, M.I.T.), resultando em críticas que foram respondidas por Searle em seu próprio texto; anos depois, autores conceituados como Jack Copeland (2007), Georges Rey (2007), Roger Penrose (2007), Stevan Harnard (2007), Mark Bishop (2007), dentre outros, abordaram o texto de Searle e elaboraram críticas para este, tanto de cunho lógico-formal quanto com relação à compreensão de preceitos básicos ligados à computação e às ciências cognitivas, o que talvez tenha causado a iniciativa do próprio Searle em publicar o seu artigo *Vinte e um anos do quarto chinês* (2007). É de grande valia mencionar a relativamente recente publicação da Oxford University Press de uma coletânea de vários desses artigos, intitulada *Views into the chinese room* (2007), que reúne grandes nomes da área da Filosofia Analítica e da Mente, a saber, Jack Copeland, Georges Rey, Roger Penrose, Stevan Harnard, Mark Bishop, dentre outros, além do próprio Searle. Neste texto, Searle clarifica ainda mais a sua argumentação original e trata de algumas críticas a esta, que surgiram desde a publicação do *Minds, brains and programs*.

O posicionamento de Searle frente ao advento da Inteligência Artificial (ou ao seu primeiro suspiro, com Alan Turing e Alonzo Church), rendeu até hoje 34 anos de discussão em torno do seu famoso argumento do quarto chinês. Dentre as várias críticas ao texto de Searle que podem ser listadas, as mais frequentes são da má compreensão da proposta original da Inteligência Artificial e dos problemas de cunho lógico que podem ser identificados no decorrer de sua argumentação. A exemplo disto, pode-se citar o texto de Harnad (2007), em que a principal crítica encontrada é a de que Searle provavelmente compreendeu mal as pretensões e o texto original de Alan Turing (*Computer machinery and intelligence* (1950)), acreditando que este tencionaria determinar de forma definitiva, através de algo chamado “teste de

Turing", se uma máquina possui ou não estados mentais. Por outro lado, Copeland (2007) ataca o texto de Searle em seu aspecto mais formal, desenvolvendo sua crítica de um ponto de vista estritamente lógico: ele aponta falácias de composição, de petição de princípio, argumentos logicamente inválidos, além de acusar Searle de cometer o mesmo equívoco que ele acusa os seus críticos de terem cometido, a saber, confundir simulação com duplicação.

Diante da macrovisão que foi traçada até agora do tema proposto, é possível, finalmente, apresentar o problema central que motivará esta pesquisa, a saber, a tentativa de responder à seguinte questão: *qual a contribuição de Searle para a filosofia da mente, no que diz respeito às discussões em torno da Inteligência Artificial?* Naturalmente, subsequentemente a esta questão central desenvolvem-se algumas questões secundárias: 1) quais as principais críticas formuladas ao texto searleano?; 2) quais dessas críticas são fortes o suficiente para invalidar algum aspecto da argumentação de Searle?; 3) o que pode ser considerado positivo na argumentação de Searle, diante de todas as críticas que o precederam? Buscaremos responder essas questões no decorrer de nosso texto.

Dessa forma, é possível apresentar agora a composição deste texto: o nosso primeiro capítulo será dedicado a apresentar os conceitos-chaves que serão usados no decorrer do restante do texto, buscando contextualizar o tema e deixar claro quais são os problemas, teorias e definições que Searle buscará criticar com seu argumento. Esse primeiro capítulo preparará o leitor para o capítulo seguinte, que por sua vez terá como função apresentar detalhadamente o argumento do quarto chinês e apontar quais são os seus "alvos", ou seja, o que o argumento pretende atingir, quais teorias pretende refutar. Após o argumento ser devidamente apresentado e esmiuçado no capítulo 2, teremos o capítulo 3 para apresentar as críticas a ele; nesse momento, serão apresentadas não apenas as críticas que Searle já acrescenta ao *Minds, brains and programs* (1980), como também críticas posteriores ao seu texto. Feito isto, o nosso último capítulo trará o desfecho que carregará a resposta da nossa pergunta metodológica, apresentada anteriormente: *qual a contribuição de Searle para a filosofia da mente, no que diz respeito às discussões em torno da Inteligência Artificial?*

O leitor perceberá que determinar exatamente que contribuição seria essa não é uma tarefa simples ou óbvia, já que muitos criticaram os posicionamentos de Searle, como veremos no capítulo 3, e nem todas as críticas puderam ser facilmente

respondidas. Não obstante, compreendendo que o pilar que sustenta toda a discussão em torno de uma inteligência artificial é o conceito de mente, ou consciência, buscaremos através desse viés demonstrar qual foi a colaboração de Searle a Filosofia da Mente, relativa a uma área desta última que tenciona não apenas desvendar o problema mente-corpo (ou mente-cérebro), como também propor um novo elemento: uma inteligência artificial.

1. CONTEXTUALIZANDO: A INTELIGÊNCIA ARTIFICIAL E ALGUMAS TEORIAS DA MENTE

Antes que possamos apresentar e discutir a esfera da problematização searleana, muito bem ilustrada pela metáfora do quarto chinês, é preciso que voltemos um pouco, resgatando as origens da disciplina ou, como muitos preferem, da engenharia ou ciência, que é a Inteligência Artificial. Além disso, nos empenharemos aqui em fornecer ao leitor o instrumental necessário para seguirmos em frente, posteriormente[3], analisando os alvos do AQC (a partir de então nos referiremos dessa forma ao Argumento do Quarto Chinês), a saber, quais teorias ele busca refutar ou invalidar, bem como as objeções principais ao argumento. Para isso, deveremos tratar neste capítulo dos conceitos de Inteligência Artificial, behaviorismo, funcionalismo e conexionismo.

1.1. Inteligência Artificial e Alan Turing

Dificilmente o advento que foi a idealização da chamada Máquina de Turing pode ser ignorado por pesquisadores da área ou mesmo por leigos desavisados: muitas vezes agraciado com o título de pai da computação, Alan Turing sobressaiu-se principalmente por cinco feitos, fortemente relacionados entre si. São eles: a resposta justificada e comprovada ao *problema de decisão* do matemático alemão David Hilbert, a idealização da sua Máquina de Turing, seguida da Máquina de Turing Universal, a elaboração da *Tese de Church-Turing* (tese desenvolvida junto ao matemático Alonzo Church) e do Teste de Turing. Apesar de todos esses feitos serem importantes para um estudo detalhado sobre a contribuição geral de Turing, em nosso caso poderemos deixar a Tese de Church-Turing à parte de nossa apresentação neste capítulo, já que Searle buscará criticar principalmente o Teste de Turing e as suas máquinas (Máquina de Turing e a Máquina de Turing Universal). Vejamos o que pode ser dito sobre isso.

[3] Faremos isso nos capítulos 2 e 3, principalmente, onde apresentaremos o AQC e em seguida as suas críticas/objeções, nessa sequência.

Em 1936, antes mesmo de se doutorar, Turing lança o seu artigo [4] *On computable numbers, with an application to the Entscheidungsproblem*[5], inspirado por um dos problemas lançados por Hilbert, conhecido como *o décimo problema de Hilbert*[6], ou o *problema de decisão*. Um problema de decisão é todo aquele que requer uma resposta positiva ou negativa (sim ou não), como por exemplo: "dados dois números *x* e *y, x* é divisível por *y*? Dessa forma, espera-se uma resposta decisiva que seja negativa ou positiva. Caso não seja possível responder, o problema seria *indecidível.* Hilbert apresentou o seu décimo problema da seguinte maneira:

> 10. Determinação da solubilidade de uma equação Diophantina.
> Dada uma equação Diophantina com qualquer número de quantidades indeterminadas e com coeficientes integrais racionais: Conceber um processo conforme o qual pode ser determinado, em um número finito de operações, se a equação é solúvel nos inteiros racionais[7]. (HILBERT, 1902, p. 421)

O problema de Hilbert, em linhas gerais, consistia em buscar determinar se uma equação seria solúvel ou não, em inteiros racionais. Ou melhor: se seria possível determinar se esse problema era decidível ou indecidível. Turing, a fim de solucionar tal problema, buscou criar um algoritmo capaz de indicar se esse problema era decidível e se sim, se a resposta que se buscava seria negativa ou positiva. Em prol de tal objetivo, buscando demonstrar que tal resposta é negativa (não, não é possível que a dada equação seja solúvel nos inteiros racionais),Turing lança a sua tese (conhecida mundialmente como *Tese de Turing*), trazendo com esta a primeira teoria concreta em computação: o *conceito* de Máquina de Turing.

Faz-se mister frisar que a Máquina de Turing corresponde a um modelo teórico em lugar de um dispositivo, posto que qualquer objeto ou qualquer pessoa pode ser tida como uma Máquina de Turing, contanto que desempenhe a função de

---

[4]TURING, Allan M. On computable numbers, with an application the the Entscheidungsproblem. In: _____. *Proceedings of the London Mathematical Society*. Series 2. Vol. 42. 1936. p. 65-230.

[5]Leia-se aqui "problema de decisão".

[6] Hilbert apresenta o seu décimo problema em 1900, numa palestra no *Congresso internacional de matemáticos* em Paris, enviando o seu artigo intitulado *Mathematical problems* um pouco antes de sua apresentação e publicado originalmente na revista alemã *Göttinger Nachrichten*. Entretanto, estaremos usando a referência mais tardia, de 1902, na revista *Bulletin of the American Mathematical Society*.

[7] "10. Determination of the solvability of a diophantine equation.
Given a diophantine equation with any number of unknown quantities and with rational integral numerical coecients: To devise a process according to which it can be determined by a finite number of operations whether the equation is solvable in rational integers" (HILBERT, 1902, p. 421).

uma. É importante afirmar que para Turing, tal "máquina" não seria nada mais do que uma pessoa executando um cálculo, computando. Quando Turing se refere a "computadores", em seu artigo de 1950, ele na verdade se refere a *humanos que computam* ou, como verificamos em seu texto (TURING, 1950, p. 433), "computadores humanos". Como se daria tal computação? Através da escrita de símbolos de um alfabeto finito em uma fita de papel uniformemente dividida em quadros. Nesse momento, o que importa são dois fatores: os símbolos que estão sendo observados e o "estado mental" daquele que computa, naquele momento. De acordo com Turing[8], "Nós conhecemos o estado do sistema se conhecermos a sequência de símbolos na fita, os quais estão sendo observados pelo computador [...] e o estado mental do computador[9]" (TURING *apud* PRESTON, 2007, p. 3). Em seguida, Turing proporá que se evite o termo "estado mental", supondo que a pessoa que opera a máquina escreva uma *nota de instrução*, explicando exatamente como a computação deve ser continuada. Tais "notas de instrução" seriam o que hoje conhecemos como *o programa*. Essa ideia é o que possibilita forjar a conexão entre o homem e a máquina, já que "[...] agora podemos construir uma máquina que realize o trabalho deste que computa[10]" (TURING *apud* PRESTON, 2007, p. 3). Através das explicações de Preston, busquemos agora compreender melhor o funcionamento de uma Máquina de Turing, através de sua explanação mais detalhada. A máquina consistiria em:

1) "Uma fita, indefinidamente longa em ambas as direções, na qual os símbolos são impressos, divididos entre quadros do mesmo tamanho, cada um podendo conter não mais de um símbolo por vez[11]" (PRESTON, 2007, p. 4).
2) "Uma cabeça móvel que (a) imprima símbolos discretos, desenhados de um alfabeto finito, na fita, (b) apague da fita um símbolo por vez e (c) leia ou

[8] Todas as citações extraídas de obras originais ou escritas em línguas estrangeiras, foram traduzidas por nós. A partir de então forneceremos os trechos originais em notas de rodapé.
[9] We know the state of the system if we know the sequence of symbols on the tape, which of these are observed by the computer . . . and the state of mind of the computer". (TURING *apud* PRESTON, 2007, p. 3).
[10] "[…] we may now construct a machine to do the work of this computer". (TURING *apud* PRESTON, 2007, p. 3)
[11] A tape, indefinitely long in both directions, on which symbols are printed, divided into square frames of the same size, each of which can contain at any one time not more than one symbol". (PRESTON, 2007, p. 4)

identifique os conteúdos de cada quadro da fita, um quadro por vez[12]" (PRESTON, 2007, p. 4).

Neste caso, o essencial é o *funcionamento*: qualquer coisa que possa *operar como* uma Máquina de Turing *será* uma Máquina de Turing. Turing não estancou os seus progressos até este ponto, ganhando enfoque pela originalidade do que chamou *Máquina de Turing Universal*, o que nos leva ao terceiro feito que engrandeceu Allan Turing. Sendo uma de suas maiores conquistas, o Teorema de Turing apresentava ao mundo o conceito de Máquina de Turing Universal. O que seria isto? Uma máquina que possuísse inscrito em sua fita a descrição codificada de uma outra Máquina de Turing, para que pudesse funcionar realizando o trabalho de outra máquina semelhante a ela. Inserir o programa de uma máquina na memória de outra consistiria em um passo fundamental para o desenvolvimento de nossos computadores/máquinas contemporâneas. Turing apresentou a Máquina de Turing e a Máquina de Turing Universal em 1936 e 1937, escrevendo só depois de alguns anos, em 1950[13], o artigo que apresentaria o que ele inicialmente chamou de *jogo de imitação*, que atualmente conhecemos como o *Teste de Turing*.

Um dos alvos do Argumento do Quarto Chinês, como veremos nos próximos capítulos, é o Teste de Turing (TT, a partir de agora em diante); a motivação por trás disto é o fato do TT simbolizar uma das primeiras demonstrações de IA Forte. Neste primeiro momento, será suficiente afirmar, preliminarmente, que Inteligência Artificial (IA, a partir de agora) Forte é o termo designado por Searle para denominar aquele ramo da Inteligência Artificial detentor da crença de que é possível duplicar a mente humana a partir de um modelo computacional. Antes de tudo, é preciso deixar claro que Alan Turing não afirma literalmente que o TT é a prova cabal de que computadores podem duplicar a mente humana; afirmamos e ressaltamos a correlação de Turing com a IA Forte pelas abordagens de filósofos como Searle, Preston e outros pesquisadores com relação a este quesito.

O Teste de Turing foi intitulado originalmente pelo seu criador de *jogo de imitação*, em seu artigo de 1950, *Computing machinery and intelligence*. Segundo

[12] A movable head that (a) prints discrete symbols, drawn from a finite alphabet, onto the tape, (b) erases one symbol at a time from the tape, and (c) reads, or identifies the contentes of each square of the tape, one square at a time". (PRESTON, 2007, p. 4)
[13] Cf. TURING, A. M. Computing Machinery and Intelligence. *Mind*, 1950.

Preston, “Turing e seus defensores insistiram que se uma máquina não pode ser distinguida de um ser humano sob estas condições, deveremos dar a ela o crédito de possuir inteligência[14]” (2007, p. 7). Que condições seriam essas? Condições sob as quais um ser humano, desempenhando o papel de um interrogador, lançaria questões a um indivíduo A e um indivíduo B, esperando respostas apenas como “sim” e “não”, a fim de descobrir o gênero de ambos os indivíduos; o que o interrogador não saberia é que um dos indivíduos não seria humano, mas uma máquina. Caso não fosse possível descobrir a não-organicidade de um dos indivíduos e o interrogador fosse enganado pela máquina, se diria que a máquina passou no teste da imitação, por ter imitado com louvor o comportamento de um ser humano pensante. A conclusão de se conferir mentalidade a uma máquina após esta ser bem sucedida no TT carrega um *background* fortemente behaviorista, o que nos leva à nossa próxima sessão.

1.2. O behaviorismo psicológico

O behaviorismo psicológico é aquela teoria segundo a qual uma teoria só seria propriamente científica caso se ativesse a características comportamentais observáveis, como o movimento do corpo. Na filosofia, o behaviorismo é aquela visão de que fenômenos mentais ou psicológicos podem ser completamente reduzidos e/ou explicados a partir de suas características físicas e movimentos observáveis. Analogamente à linguagem computacional, poderíamos observar e analisar estados mentais partindo apenas dos *inputs* e *outputs* relacionados ao indivíduo estudado. O behaviorismo psicológico, bem como o filosófico, são a porta de entrada para uma outra corrente de pensamento que atualmente é a mais popular dentro da Filosofia da Mente: o *funcionalismo*.

1.3. O funcionalismo de Hilary Putnam

Para compreendermos a natureza do funcionalismo, precisamos antes entender como se deu a transição do pensamento pautado numa teoria behaviorista até aquele pautado no que hoje chamamos de funcionalismo. Essa transição se deu

[14] “Turing and his defenders then insist that if a machine cannot be distinguished from a human being under these conditions we must credit it with intelligence”. (PRESTON, 2007, p. 7)

por ter-se identificado duas falhas principais no behaviorismo: ignorar ou simplesmente negar os aspectos internos de estados mentais, como a dor, por exemplo. A dor possui aspectos *qualitativos* (que em Filosofia da Mente chamamos de *qualia*) que não podem ser deixados à parte: sentir dor não é apenas mover-se como uma pessoa que sente dor, gemer como quem sente dor ou implorar por remédios analgésicos – o sentimento da dor se revela principalmente na introspecção, num sentimento que acomete o âmago (altamente privado) de cada um. A segunda falha do behaviorismo se deu ao tentar decompor estados mentais complexos a uma explicação comportamental altamente reducionista. Por exemplo, o estado mental referente ao desejo de tirar férias no Caribe é circundado por várias sentenças de natureza condicional: “se estou cansado de ficar no Havaí, então devo ir ao Caribe”, “se hoje é sexta, e sextas geralmente implicam início de uma folga, então eu deveria viajar ao Caribe na minha folga”, “tenho duas opções de viagem onde o Caribe é a mais barata entre elas; já que eu tenho pouco dinheiro, então devo viajar para o Caribe e não optar pelo outro destino” etc. O behaviorismo não deu conta de análises meramente comportamentais desse tipo de estado mental, e por isso muitos abandonaram essa corrente de pensamento.

O funcionalismo, por sua vez, seria uma promessa de sanar os problemas carregados pelo behaviorismo, trazendo características inovadoras. Enquanto o behaviorista se preocuparia apenas com o comportamento, ou com a relação entre o sujeito e o mundo que o cerca, o funcionalista se preocupa em compreender os estados mentais a partir das suas relações causais não apenas com o ambiente “externo”, mas também com relação a outros estados mentais. Para o funcionalista, os estados mentais não seriam apenas o resultado de uma relação de *inputs* e *outputs* entre o meio ambiente e o sujeito pensante – muito mais estaria em jogo. O behaviorista comete o erro de *reduzir* o mental a entradas e saídas de dados, e é justamente isso que o funcionalista buscará evitar.

Poderíamos tentar diferenciar o funcionalismo do behaviorismo também da seguinte maneira: enquanto o behaviorista se preocupa com o comportamento, o funcionalista se preocupa com o funcionamento. Mas de que forma? Podemos ilustrar isso com o seguinte exemplo. Imaginemos que em outro planeta exista um alienígena cuja composição corporal não é baseada no carbono, como é o caso dos humanos. Vamos supor que seja baseado em, digamos, silício. Entretanto, o alienígena ainda assim teria órgãos muito similares aos nossos: um órgão

responsável pela sua respiração, como os nossos pulmões, outro por bombear o sangue, outro para filtrá-lo, e por aí vai. Mesmo que seu corpo não seja idêntico ao nosso no que concerne à sua composição molecular, ele *funciona* de forma idêntica ao nosso. Para o funcionalista, a *funcionalidade* isomórfica (o que acabamos de descrever) já bastaria para tomar o alienígena como objeto de estudo do funcionamento do nosso corpo, já que existe a tal funcionalidade isomórfica entre as espécies. O mesmo pode ser aplicado às mentes, de acordo com o funcionalista. Em suma, para ele, "O que é importante para a existência de uma mente não é a matéria da qual a criatura é feita, mas a estrutura das atividades internas mantidas por essa matéria" (CHURCHLAND, 2004, p. 69).

Como podemos detalhar melhor a aplicação do funcionalismo à esfera do mental? Partindo do início: o pai do funcionalismo foi o filósofo americano Hilary Putnam, cujas crenças se baseiam na premissa de que o mental seria mais uma questão de funcionamento do que de substância; tal assertiva carrega uma conclusão quase que imediata: se o que importa no mental não é a substância mas apenas o funcionamento, a partir do momento que algo *funciona* da mesma forma que uma mente, então é possível afirmar que não apenas a mente foi simulada, mas fundamentalmente *duplicada*. Putnam, como poderíamos esperar, relacionou a mente humana com uma Máquina de Turing, propondo a sua analogia; entretanto, não parou por aí: em 1967, Putnam[15] afirma que os humanos *são* o que ele mesmo chamou de "autômato probabilístico". O que seria isto? O autômato probabilístico seria algo similar à Máquina de Turing, diferindo apenas no sentido de que é permitido que as transições entre os estados da máquina tenham diversas probabilidades, em vez de serem deterministas. No autômato, poderíamos enviar *inputs sensoriais* para receber *outputs motores*, e neste momento o leitor poderá se lembrar do behaviorismo já mencionado anteriormente. Em contrapartida, não se deve esquecer que o funcionalismo aqui abandona as falhas do behaviorismo, se preocupando com o que este deixou para trás: o autômato probabilístico, por lidar com probabilidades, amplia bastante o leque de *outputs* que poderiam ser implicados diante de um único *input*, o que faz com que Putnam facilmente o associe a um ser humano.

[15] Cf. PUTNAM, H. The nature of mental states. In: _____. *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 1967.

Ainda que dotado de múltiplas probabilidades referentes a *inputs* e *outputs*, os funcionalistas reconheceram que o autômato probabilístico de Putnam ainda não conseguiria atingir o nível de complexidade humano, tendo em vista que este não carregaria mais do que apenas um estado (nesse caso, psicológico), não podendo desempenhar muito mais do que uma operação (nesse caso, mental). Esta ideia então foi deixada de lado e um novo tipo de funcionalismo tornou-se mais dominante: aquele no qual tem-se em vista que os estados mentais não podem ser individualizados apenas por estados de uma máquina, mas através de suas conexões com *inputs* sensoriais, entre si mesmos e com *outputs* behaviorísticos. Esse novo tipo de funcionalismo é agora o mais evidenciado na Filosofia da Mente e se encontra no núcleo da ciência cognitiva.

1.4. O Computacionalismo de Jerry Fodor

Fodor é considerado o pai da teoria computacionalista, que é, para dizer o mínimo, bastante polêmica. Searle[16], juntamente com Dreyfus[17], foi um dos filósofos que mais criticou o ponto de vista de Fodor, bem como toda a proposta do que ele chamou de IA Forte. A fim de evitar enganos na nossa explanação, vamos tomar separadamente duas visões distintas do Computacionalismo: em primeiro lugar, analisemos o que Preston dirá a respeito disto, posto que este autor se comprometeu a expor em algumas páginas a trajetória das ciências cognitivas, da inteligência artificial e do próprio Argumento do Quarto Chinês searleano, que veremos no próximo capítulo. A visão de Preston será interessante para nós nesse momento por tender mais à imparcialidade. Seguiremos apresentando a visão de Searle com relação ao computacionalismo, apresentando, por último, o discurso de Fodor, pai da teoria computacionalista, buscando esclarecer o que de fato ele postula.

À luz do discurso de Preston, podemos destacar duas características principais da teoria computacionalista de Fodor: 1) pensamento racional é apenas

[16]A criíica mais forte de Searle se encontra em seu artigo *Minds, brains and programs* (1980), mas também pode ser encontrada em *Minds and brains without programs* (1987), *Mind* (2004) e *Minds, brains and science* (1984).

[17]Pode-se destacar aqui uma das obras mais famosas de Dreyfus, a qual muitos pesquisadores sempre recorrem ao tratar da sua crítica à Inteligência Artificial: *What computers can't do* (1972); posteriormente, Dreyfus ainda publica o texto *What computers still can't do* (1992).

uma questão de preservação da verdade em certas inferências e 2) para o computacionalismo, a cognição humana se trataria apenas do cérebro processando símbolos unicamente de forma sintática, onde as características semânticas seriam irrelevantes. A visão de Searle com relação ao computacionalismo não difere muito disto: para Searle, o computacionalista acredita que a mente é apenas um programa de computador sendo executado no cérebro, que seria uma espécie de hardware biológico; dessa forma, o TT seria decisivo para atestar se um dispositivo possui mentalidade/consciência ou não.

Ao analisarmos, logo em seguida, o discurso de Fodor, poderemos concluir muito mais do que foi dito acima, nos encontrando em tal situação que seria possível rever a visão de Searle com relação ao computacionalismo. De acordo com Fodor, a Inteligência Artificial seria basicamente uma *engenharia*, e não propriamente uma *ciência*. Tomemos como exemplo, assim como Fodor o fez, a ciência cognitiva: ela busca *compreender* o pensamento, assim como outras ciências buscam compreender os seus objetos de estudo. A Inteligência Artificial, por sua vez, busca apenas a construção de artefatos inteligentes, tendo em vista *simular* a inteligência. Dessa maneira, chegamos então ao que pode ser considerado fulcral na abordagem de Fodor: a questão da simulação. Fodor é bastante direto ao afirmar, com relação à simulação, que ele não vê interesse científico em simular o comportamento, que não acredita que o objetivo da ciência cognitiva seja esse, além de afirmar literalmente que seria loucura tentar seriamente simular o comportamento inteligente. Para compreender a motivação de afirmações tão fortes, perscrutemos quais seriam as implicações da simulação de um comportamento inteligente, para Fodor.

A simulação seria desnecessária para uma investigação científica por não significar nenhuma prova de que uma teoria seja válida ou não. Não é preciso simular um fenômeno físico para compreender como ele funciona, por exemplo. Além disso, muitos fenômenos sequer podem ser simulados, e isso não faz com que uma área de pesquisa (daquele fenômeno) mereça descrédito. Tendo isso em vista, simular uma *parte* do comportamento inteligente, como no caso do TT, seria apenas um interesse marginal. É preciso, alerta Fodor, compreender o limiar entre a ciência cognitiva e Inteligência Artificial: enquanto esta segunda apenas se preocupa em compreender a inteligência através de simulações, a primeira busca compreendê-la através de suas interações com outros processos mentais e tendo isso em vista, pode visar construir um ambiente estrategicamente desenvolvido para que se possa

estudar tais processos um a um. “Deixando de lado a engenharia, além do fato de que seria legal fazer robôs e assim por diante, eu não vejo interesse científico em simular o comportamento[18]”, afirma conclusivamente Fodor (1997, p. 87).

Fodor ainda ressalta características importantes sobre o Teste de Turing: tendo em vista que a simulação é extremamente periférica e de pouca importância para a compreensão da mente, o TT deixa de ser tão impressionante, nos possibilitando uma visão mais cética em torno dele. O TT se trata de atestar a inteligência em um dispositivo mecânico, como vimos até então; isso se dá através da demonstração de um comportamento similar ao comportamento humano quando sob as mesmas circunstâncias. Não obstante, não podemos esquecer que o interrogador, incumbido de perscrutar e descobrir se um dos indivíduos ocultos é ou não um ser humano, jamais fora avisado de que essa seria a sua tarefa: ele deveria apenas determinar o sexo dos indivíduos, e não a sua natureza – orgânica ou não. Dessa forma, Fodor questiona: é possível que o interrogador fora enganado pela máquina apenas por não possuir o conjunto de perguntas corretas? Se ele soubesse que um dos indivíduos era uma máquina, ele não seria mais perspicaz em descobrir qual dos dois seria? É interessante deixar claro nesse momento que, ainda que Fodor levante esse questionamento, apenas a versão original do teste é dessa forma, mas em outras versões, o interrogador é ciente de que um dos participantes é uma máquina buscando se passar por um ser humano.

Obviamente, a discussão aqui não deve seguir um rumo distinto ao que estamos nos propondo: não buscamos verificar quais perguntas seriam as *corretas* para determinar a humanidade de um humano ou a artificialidade de uma máquina mas, diferentemente disso, iluminar o Teste de Turing a partir da perspectiva cética de que da mesma forma que ao passar no TT, uma máquina poderia ser considerada inteligente, assim também o interrogador poderia simplesmente não ter feito as perguntas corretas.

Quando questionado a respeito da metodologia correta para se estudar a mente, Fodor (1997, p. 88) simplesmente responde que esta seria a mesma utilizada na Geologia ou em qualquer outra ciência:

---

[18] “Aside from engineering, aside from the fact that it would be nice to make robots and soo n, I don’t see the scientific interest of simulating behavior”. (FODOR, 1997, p. 87).

> Você desenvolveria teorias que explicam o comportamento observável. Então você constrói um ambiente artificial no qual os detalhes da teoria podem ser testados. Eu não acho que haja alguma resposta interessante à questão do que nós devemos fazer na ciência cognitiva e que pudesse ser diferente da resposta à questão do que nós deveríamos fazer em geologia[19].

Por último, mas não menos importante, é preciso dar enfoque ao problema da intencionalidade, tão bem exposto por Searle, ao qual Fodor responde da seguinte maneira: o principal questionamento a respeito da intencionalidade é "o que é necessário para um estado físico possuir propriedades semânticas?[20]" (FODOR, 1997, p. 89). A este questionamento o TT não é endereçado, já que o teste não está interessado em descobrir se um aparelho eletrônico possui propriedades semânticas. Então, qual questionamento o TT se destina a responder? Este seria: "posto que se tem um dispositivo cujos estados possuem propriedades intencionais, qual tipo de dispositivo seria tal que suas transições de estado a estado fossem inteligentes?[21]" (FODOR, 1997, p. 89). Notemos aqui a distinção fundamental entre os questionamentos levantados: enquanto Searle se preocupa com a *intencionalidade*, como vimos na nossa introdução, Turing se preocupa com a *inteligência*; essa mesma preocupação, ou escopo, de acordo com Fodor, também é relacionada à disciplina da Inteligência Artificial e à ciência cognitiva como um todo, como explicita Fodor (1997, p. 90):

> A principal ideia é explorar o fato de que propriedades sintáticas podem preservar propriedades semânticas. Esta é a ideia básica de uma teoria demonstrativa. Searle está perfeitamente certo quando diz que o empreendimento de Turing não lhe diz o que é preciso para um estado possuir propriedades semânticas. Seja qual for a teoria correta da intencionalidade, não é o que a ciência cognitiva está em busca. Eu penso que você pode ter uma teoria da intencionalidade, mas não é isso o que a ciência cognitiva está tentando prover. Se presumirmos a intencionalidade, a questão se torna: podemos ter uma teoria sintática da inteligência?[22]

---

[19] "You would develop theories that explain the observable behavior. Then you construct artificial environments in which details of the theory can be tested. I don't think there is any interesting answer to the question of what we should do in cognitive science that would be different from the answer to the question of what we should do in geology". (FODOR, 1997, p. 88)

[20] "What is it for a physical state to have semantic properties?" (FODOR, 1997, p. 89)

[21] Given that you have a device whose states have intentional properties, what kind of device would be such that its state transitions are intelligent?" (FODOR, 1997, p. 89).

[22] "The main idea to exploit the fact that syntactical properties can preserve semantic properties. That is the basic idea of proof theory. Searle is perfectly right when he says that Turing enterprise does not tell you what it is for a state to have semantic properties. Whatever the right theory about intentionality is, it is not what cognitive science is going on about. I think you can have a theory of intentionality, but

Em termos de crítica, Dreyfus estabelece uma postura mais radical que a de Searle (como vimos rapidamente em nossa introdução), afirmando que é *impossível* mecanizar a inteligência. Tal afirmativa é recorrente do fato de ainda não se ter registro de uma simulação perfeita do comportamento racional humano. De volta ao problema da simulação, como já aqui dissemos, a simulação não é a prova cabal de que uma teoria é válida ou que deve ser reconhecida como tal; uma ciência pode ser considerada positiva (no sentido de produzir informações válidas) se lhe é possível estabelecer um conjunto de leis tendo como base a observação de um dado fenômeno e se é possível testar suas teorias em um ambiente experimental, refazendo eventos e prevendo outros. Por ora, deixemos de lado o computacionalismo até que possamos nos aprofundar na abordagem da crítica de Searle.

1.5. Inteligência Artificial e suas bifurcações

A Inteligência Artificial nasce em meados dos anos 1950 juntamente com os textos mais importantes de Turing, como já mencionamos aqui; entretanto, Margaret Boden traça o seu início já em 1943, identificando o artigo de Warren McCulloch e Walter Pitts[23], onde foi proposta uma correspondência entre as relações psicológicas entre os neurônios e as relações lógicas entre as proposições. A partir de então, podemos identificar duas bifurcações na história da Inteligência Artificial, sendo uma delas *histórica* e a outra *qualitativa*. Esta última, como Searle bem poderá nos apresentar, seria da distinção entre Inteligência Artificial Forte e Inteligência Artificial Fraca. Searle nos diz que os pesquisadores da IA Fraca consideram o computador como uma ferramenta poderosa para o estudo da mente humana, não sendo este nada além disso. Já no caso da IA Forte, as pretensões de seus defensores seriam mais ambiciosas, interpretando o computador não como uma ferramenta de estudo da mente, mas como o objeto de estudo ele próprio, já que, segundo seus postuladores, seria possível duplicar a mentalidade humana ao ponto de não haver necessidade de estudar o cérebro de um homem quando um computador estivesse

---

this is not what cognitive science is trying to give. If we assume intentionality, the question becomes, Can we have a syntactical theory of intelligence?" (FODOR, 1997, p. 90)

[23] Cf. McCULLOCH, W.S. e PITTS, W.H. A logical calculus of the ideas immanent in nervous activity. In: _____. *Bulletin of mathematical biophysics*. 1990. p.33-115.

disponível. Nós já vimos na seção anterior a versão de Fodor a respeito dessa visão de Searle, então seguiremos adiante, tendo em vista a bifurcação história da qual falamos anteriormente.

Segundo Preston (1997, p. 11), a IA segue em dois caminhos ou programas separados, sendo um deles o que ele chama de *clássico* e o outro de *conexionista*. Inicialmente, temos o programa clássico, influenciado mais pelo legado de McCulloch do que de Turing; esse programa voltava-se ao desenvolvimento de programas desenhados para provar teoremas de áreas da matemática, como a lógica, álgebra e geometria, assim como jogos, como podemos ver a partir de Preston (2007, p. 11-12):

> Programas desenvolvidos para provar teoremas em áreas da matemática tais como a lógica, geometria e álgebra estavam lado a lado de programas de jogos, dedicados a damas, xadrez ou jogos de cartas, e 'resolução de problemas', programas que abordam enigmas intelectuais, tais como 'torres de Hanoi', 'pontes de Konigsberg', 'o vendedor viajante', 'missionário e canibais', etc[24].

De acordo com Feigenbaum & Feldman (*apud* PRESTON, 2007, p. 12), a utilização dos jogos se dava pelo fato de ambientes de jogos são muito úteis para estudar a natureza e estrutura de processos complexos de resolução de problemas, já que através dos jogos podemos resolver charadas, quebra-cabeças e outras atividades que exigem inteligência e raciocínio.

Os programas mencionados surgiram em paralelo, mas a ênfase foi deslocada para um deles depois de certo tempo, apesar do outro não ter sido abandonado: as luzes se voltavam para o programa conexionista que buscava estudar temas como a compreensão da linguagem natural, aprendizado, aprendizado de máquina, sistemas inteligentes, dentre outros, apesar desse tipo de pesquisa não ser exclusivo desse programa. Para que essa pesquisa fosse dada cabo, foram desenvolvidas as *redes neurais*; como o termo sugere, tal recurso se encarregaria de simular artificialmente as redes neurais do cérebro humano, através de neurônios artificiais (processadores, por exemplo) com certa capacidade de processamento, conectados uns aos outros, formando uma espécie de teia ou rede,

[24] "Programs designed to prove theorems in areas of mathematics such as logic, geometry, and algebra rubbed shoulders with games-playing programs, devoted to draught (checkers), chess, or card games such as bridge, and 'problem-solvers', programs which would address intellectual puzzles such as the 'towers of Hanoi', the 'bridges of Konigsberg', the 'travelling salesman', 'missionary and cannibals' cases, etc. (PRESTON, 2007, p. 11-12).

onde cada unidade de processamento é conectada e se conecta a outra unidade, mimetizando a organização neural do cérebro.

A primeira simulação de uma rede neural foi desenvolvida pelo Instituto de Tecnologia de Massachusetts em 1954, e visava basicamente o reconhecimento e preenchimento de padrões. Segundo Preston (1997, p.13), “As redes neurais foram ditas tendo várias similaridades com o cérebro humano, no que diz respeito à estrutura e operação. Mas isso não quer dizer que os primeiros sejam modelos precisos deste último”. Para Fodor, redes neurais não explicam praticamente nada sobre o aprendizado, já que se houvesse alguma explicação, essa seria a respeito do aprendizado como processo de inferência estatística, o que não é o caso. Com tais dispositivos, não se teria uma nova teoria da aprendizagem, apenas uma teoria *estatística* padrão da aprendizagem. Eis as palavras do filósofo, a respeito das redes neurais,

> Pensando nelas como dispositivos de aprendizagem, eu penso que elas são pacotes analógicos de estatísticas. Para obter algum sucesso com esses dispositivos, mostrar que esta é a arquitetura que, por si só, está levando a novas descobertas e a uma nova compreensão, é fazer algo que não se pode fazer com uma análise estatística. Eu não conheço nenhuma demonstração onde este é o caso[25]. (FODOR, 1997, p. 95)

Teorias da mente como o funcionalismo e o computacionalismo, além do programa conexionista, os quais tratamos brevemente nas sessões anteriores, formam o conteúdo da bifurcação qualitativa à qual nos referíamos anteriormente. Como evidenciou Preston (1997, p.15), foi suposto que a IA Forte simplesmente não poderia ser verificada na prática, que esse tipo de postura ou postulado não poderia ser evidenciado; em defesa de Searle (neste aspecto), Preston menciona vários autores defensores de uma postura mais ambiciosa relativa ao desenvolvimento artificial do comportamento psicológico da mente humana, a saber, Dennett (1985), Simon (1980) e Newell (1958). Em suma, “Se você acha que não conhece ninguém que endosse a IA Forte, você não olhou bem o suficiente![26]” (PRESTON, 1997,

[25] “Thinking of them as learning devices, I think they are analog statistics packages. To get some success with these devices, to show that it is the architecture itself tha is leading to new discoveries and a new understanding, is to do something that you cannot do with statistical analysis. I don’t know any demonstration where this is the case”. (FODOR, 1997, p. 95).

[26] “If you think you don’t know of anyone who endorses Strong AI, you haven’t looked hard enough!”. (PRESTON, 1997, p. 16).

p.16). Por ter encontrado defensores da IA Forte, Searle encontrou-se motivado a elaborar o seu AQC, como veremos em nosso próximo capítulo.

2. O ARGUMENTO DO QUARTO CHINÊS

Até então, desenhamos um quadro geral do nosso tema, apresentando e esmiuçando algumas linhas de pensamento que permeiam a Filosofia da Mente e discutindo sobre a Inteligência Artificial, terreno tão fértil para a obra searleana. Feito isto, chega o momento de recorrermos ao coração de nossa abordagem, o AQC dando ênfase não apenas a tal experiência de pensamento, mas buscando sinalizar exatamente os alvos (ou seja, para onde a crítica é apontada) que Searle busca atingir através do AQC.

2.1. Natureza e objetivos do AQC

Antes de refazermos os passos de Searle, detalhando passo-a-passo o experimento do AQC, é preciso que estejamos cientes de sua natureza e ao que se propõe. Tal argumento, como é bastante comum em Filosofia da Mente, é o que chamamos de *Gedankenexperiment*, a saber, uma experiência de pensamento que, em nosso contexto particular[27], tem em vista testar alguma teoria da mente através do deslocamento da terceira para a primeira pessoa, onde aquele que busca experimentar abstratamente uma dada vivência pode ser capaz de determinar, em seguida, se tal teoria se aplica ou não à realidade de uma mente humana tendo como base a sua própria.

Um segundo aspecto da natureza do argumento em questão é a sua busca pela elucidação do conceito de *inteligência*, impulsionada pelo questionamento inicial de Alan Turing em seu artigo de 1950[28]: “As máquinas podem pensar?”. Intrínseca aos objetivos do AQC está a pressuposição de que, em resposta à pergunta de Turing, máquinas poderiam, sim, pensar; entretanto, apenas um tipo bastante particular de máquina, o cérebro humano, ou pelo menos alguma máquina que possuísse os mesmos poderes causais que ele. O pensamento, afirma Searle, jamais poderá ser atribuído a *programas*, apenas a máquinas[29].

---

[27] As experiências de pensamento não são exclusividade da Filosofia, sendo comumente encontradas na área da Física, dentre outras ciências.
[28] Cf. TURING, A. M. Computing Machinery and Intelligence. *Mind*, 1950.
[29] Tais máquinas podem ser os seres humanos ou outras animais de outras espécies que possuem intencionalidade e capacidades similares às do cérebro humano.

Dito isto, nos parece razoável analisar agora os objetivos do AQC, que seriam, basicamente, os esforços com vistas a suportar duas proposições. São as seguintes:

1) “Intencionalidade em seres humanos (e animais) é um produto de características causais do cérebro. [...] Certos processos cerebrais são suficientes para haver intencionalidade[30]”.[31] (SEARLE, 1980, p.417)

2) “Instanciar um programa de computador jamais será, por si só, uma condição suficiente para haver intencionalidade[32]”. (SEARLE, 1980, p.417)

Para sustentar a veracidade dessas premissas, a forma do AQC será tal que deverá mostrar como um ser humano poderia instanciar um programa de computador e ainda assim não possuir intencionalidade.

Como consequência dessas duas premissas, temos que:

3) “A explicação de como o cérebro produz intencionalidade não pode ser que ele o faça instanciando um programa de computador[33]”. (SEARLE, 1980, p.417).

Tomemos isso como uma consequência de 1) e 2).

4) “Qualquer mecanismo capaz de produzir intencionalidade deve ter poderes causais iguais àqueles do cérebro"[34]”. (SEARLE, 1980, p. 417)

---

[30] “Intentionality in human beings (and animals) is a product of causal features of the brain I assume this is an empirical fact about the actual causal relations between mental processes and brains. […] Certain brain processes are sufficient for intentionality”. (SEARLE, 1980, p. 417)

[31] Searle considera tal premissa verdadeira considerando que seria ela um fato empírico em torno da constituição do cérebro e das suas relações causais com o mental. Encontramos mais traços desse posicionamento searleano, nomeado *Naturalismo biológico*, em suas outras obras *Intencionalidade* (2007) e *Minds, brains and science* (1995).

[32] “Instantiating a computer program is never by itself a sufficient condition of intentionality”. (SEARLE, 1980, p. 417)

[33] “The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program”. (SEARLE, 1980, p. 417)

[34] “Any mechanism capable of producing intentionality must have causal powers equal to those of the brain”. (SEARLE, 1980, p. 417)

Essa é uma consequência trivial de 1) e explicita um posicionamento de Searle que não deve ser excetuado ou esquecido: a defesa do filósofo não se pauta na afirmação de que nenhuma máquina jamais poderá pensar, aprender, raciocinar ou possuir intencionalidade como cérebro humano; na visão de Searle, o cérebro também seria uma máquina, e qualquer máquina que possuísse os mesmos poderes causais que o cérebro, também poderia ser adjetivado da mesma maneira que este.

5) "Qualquer tentativa literal de criar intencionalidade artificialmente (IA Forte) não poderia obter sucesso apenas pelo desenvolvimento de programas, mas teria de duplicar os poderes causais do cérebro humano[35]". (SEARLE, 1980, p. 417)

   Essa inferência segue de 2) e 4).

O que traçamos até então pode ser compreendido como os *objetivos gerais* do argumento de Searle, que se voltam ao que ele considera Inteligência Artificial Forte como um todo; como *objetivos específicos*, temos a crítica diretamente voltada ao programa de computador desenvolvido por Roger Schank[36] e seus colegas em Yale (ou aos programas em geral, partindo da premissa de que programas seriam suficientes para a intencionalidade), ao computacionalismo (ou modelo computacional de mente), o cognitivismo, behaviorismo, funcionalismo e o Teste de Turing. Tendo isto em vista, é chegado o momento de perscrutarmos a literatura de Searle percorrendo de suas intenções às suas conclusões.

2.2. Compreendendo o AQC

O ponto de partida da apresentação do AQC é o programa desenvolvido por Schank, que tem como função simular a habilidade humana de compreender histórias. É característica do ser humano conseguir compreender histórias e deduzir informações adicionais sobre elas sem que essas informações lhe tenham sido

[35] Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain". (SEARLE, 1980, p. 417)
[36] Cf. SCHANK, R. C. & ABELSON, R. P. *Scripts, plans, goals, and understanding.* Lawrence Erlbaum Press: Hillsdale, 1977.

dadas previamente. O exemplo que Searle nos fornece ilustra bem o que seria tal habilidade:

> Um homem foi a um restaurante e pediu um hambúrguer. Quando o hambúrguer chegou, estava queimado e o homem saiu do restaurante com raiva, sem pagar pelo hambúrguer ou deixar uma gorjeta. Agora, lhe é perguntado '– O homem comeu o hambúrguer?', provavelmente você irá responder 'Não, ele não comeu'. Da mesma forma, se lhe é dada a seguinte história: '- Um homem foi a um restaurante e pediu um hambúrguer; quando o hambúrguer chegou, ele estava bastante satisfeito com ele; e quando saiu do restaurante, deixou uma enorme gorjeta à garçonete antes de pagar a conta', e lhe é feito o questionamento, '- O homem comeu o hambúrguer?', você presumidamente responderá, - 'Sim, ele comeu o hambúrguer'[37]. (SEARLE, 1980, p. 418)

Para simulá-la (a habilidade), seriam fornecidas várias informações ao programa, tais quais o modo de comportamento social do ser humano, nossa concepção em torno de restaurantes, qualidade, dentre outros quesitos que são necessários se conhecer para julgar se um alimento é bom ou não e se deve ser pago ou não. De posse de tais informações, seria apenas preciso que o programa cruzasse dados e interpretasse que naquela dada situação, o movimento mais previsível do ser humano, sujeito daquela história, seria agir de modo A ou B. O problema que envolve essa simulação se instaura à medida que, segundo Searle, partidários da Inteligência Artificial Forte afirmam que tal máquina (dotada do programa de compreender histórias, o programa de Schank) não apenas simularia a habilidade humana, mas a *duplicaria*, de modo que:

1) Poderia-se dizer que a máquina literalmente compreende as histórias e dá respostas às questões que lhe são fornecidas, e

2) O que é realizado pela máquina e pelo programa explica a habilidade humana de compreender histórias e responder perguntas sobre elas.

---

[37] "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip." Now, if you are asked -Did the man eat the hamburger?" you will presumably answer, ' No, he did not.' Similarly, if you are given the following story: '-A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill," and you are asked the question, -Did the man eat the hamburger?,-' you will presumably answer, -Yes, he ate the hamburger"". (SEARLE, 1980, p. 418)

Segundo Searle, há uma diferença marcante entre o que ele nomeia Inteligência Artificial Fraca e Inteligência Artificial Forte: a primeira lançaria mão de programas de computador para *compreender* habilidades e capacidades da mente humana, simulando-as num ambiente virtual; a segunda, em contrapartida, não tomaria os programas computacionais como uma mera ferramenta para estudo da mente, mas como um substituto para ela, no sentido de que seria possível duplicá-la virtualmente e não necessitaríamos mais nos limitarmos ao estudo do mental nos voltando apenas à mente do ser humano. Podemos encontrar nos tópicos 1) e 2) posições de uma IA Forte que intrigam Searle ao ponto de impelí-lo a desenvolver a experiência de pensamento do AQC, que veremos a seguir.

A experiência de pensamento se dá com a situação imaginária que aos poucos vai se formando: devemos imaginar que estamos, sozinhos, em uma sala fechada, onde nos é dado um montante de escritos em chinês. Além disso, imaginemos também que não conhecemos absolutamente nada de chinês de modo que não conseguimos diferenciar seus ideogramas dos ideogramas japoneses ou até mesmo de meros rabiscos. Em seguida, nos é fornecida uma segunda quantia de escritos em chinês, junto com um conjunto de regras sobre como relacionar o primeiro montante de escritos com o segundo, estas regras estando em português. Essas regras nos possibilitam fazer a correlação entre o primeiro grupo de escritos e o segundo apenas de maneira formal, já que não compreendemos o que tais símbolos chineses significam. Logo após nos é dado um terceiro montante de escritos em chinês em conjunto com algumas instruções, em português, que nos indicariam como correlacionar os símbolos do terceiro montante com os do primeiro e segundo, nos habilitando a fornecer de volta certos símbolos em chinês em resposta a certos tipos de símbolos que me são dados no terceiro montante de escritos em chinês. Vamos recapitular o que temos até então:

1) Uma pilha de escritos em chinês que chamaremos de A;
2) Uma segunda pilha de escritos em chinês que chamaremos de B;
3) Um conjunto de regras em português que nos ensinam a relacionar A com B. Chamemos tal conjunto de AB.
4) Uma terceira pilha de escritos em chinês que chamaremos de C;

5) Um conjunto de instruções em português que nos ensinam a correlacionar C com A e B, nos tornando capazes de algo além do que uma simples relação, nos tornando aptos a fornecer em troca alguns outros símbolos em chinês, consultando o que já dispomos. Chamemos esse conjunto de ABC.

Nesta primeira fase, temos apenas a visão da pessoa que está dentro da sala com essa quantidade de informação, em grande parte, sem sentido; a única coisa que faz sentido para nós são as instruções e regras que nos foram dadas em nossa língua mãe, que nos ensinam a relacionar um símbolo de formato X com outro de formato Y e talvez fornecer em seguida um terceiro símbolo de formato Z. Passemos então à segunda fase: enquanto isso, *fora da sala*, há pessoas que nomeiam nossas pilhas de escritos e nossas instruções de forma diferente. Na visão das pessoas de fora,

1) ‘A’ é chamado de *código* (ou *script*, de acordo com Searle);
2) ‘B’ é chamado de *história*;
3) ‘C’ é chamado de *perguntas*;
4) Aos símbolos que fornecemos para eles, através do que aprendemos com o nosso livro de instruções ABC, eles chamam de *respostas às questões*;
5) O nosso conjunto de regras que recebemos juntamente com B, o qual nomeamos AB, eles chamam de *o programa*.

Relatamos até então de maneira precisa o modo através do qual Searle apresenta o seu AQC em seu texto. Não obstante, é comum que algumas dúvidas surjam a partir das colocações searleanas em torno dos termos técnicos da área da Computação. Segundo Searle, o programa seria apenas o que chamamos aqui de AB, e A seria o script, uma espécie de código usado em Computação. Identificamos que possa haver, talvez, um exagero de elementos que compõem o AQC (Searle por vezes foi acusado de conhecer pouco a área da Computação e em alguns momentos versar livremente sobre a mesma, cometendo alguns erros, muitos deles apontados no presente argumento, o qual apresentamos[38]); desta feita, prezamos

[38] Um exemplo claro de tal crítica encontra-se no artigo de Harnad. Cf. HARNAD, Stevan. Minds, machines, and Searle 2: what's right and wrong about the chinese room argument. In: _____. *Views*

por apontar ao leitor um modo alternativo (e distinto do modo adotado por Searle) de interpretação[39], mais sucinto.

A partir daí, teríamos que A é o nosso contexto, a saber, o contexto no qual a história se insere, carregando todo o tipo de informação que versa sobre elementos socioculturais que circundam a atmosfera do que é contado; por sua vez, mantendo o que foi dito por Searle, B e C seriam a história e as questões acerca da história, respectivamente; os símbolos fornecidos de dentro para fora da sala seriam as respostas às questões e, finalmente, o programa compreenderia AB *e* ABC, posto que ambos são instruções[40].

Como o leitor já deve ter percebido, na visão de quem está fora da sala, o que acontece é que nos são fornecidas perguntas em chinês, às quais somos capazes de dar respostas em chinês. O que fazemos no interior da sala é relacionar símbolos e mais símbolos que, à propósito, não possuem nenhum significado para nós; apenas executamos uma tarefa que nos é pedida.

Dando continuidade à nossa história, passemos a uma terceira fase, onde as pessoas de fora da sala nos dão informações cem por cento em nossa língua materna e nos pedem que forneçamos respostas às perguntas elaboradas tendo como base uma certa história. Como se era de esperar, respondemos todas as perguntas com êxito. Em seguida, o que nos é pedido é que lidemos com os símbolos em chinês, novamente dando respostas (em chinês) a questões (em chinês) apenas lançando mão de nossas regras e instruções (em português). Imaginemos que nos tornamos tão habilidosos nessa tarefa que damos cabo com a mesma facilmente, fornecendo todas as respostas às perguntas ou, ao nosso ver, fornecendo bastante símbolos sem sentido depois de correlacionarmos, com a ajuda de ABC, o conteúdo de C, A e B.

Após o nosso terceiro momento, ocorre às pessoas externas à sala a ideia de que compreendemos histórias em chinês tão bem quanto compreendemos aquelas em português, já que demonstramos a habilidade de responder às perguntas com

---

into the chinese room: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

[39] Faz-se mister deixar claro que tais esclarecimentos possuem como fins facilitar a compreensão do leitor com relação ao que fora apresentado, caso sejam despertas dúvidas ou algum tipo de confusão em relação ao que já foi exposto.

[40] Mesmo em nossa versão diferenciada, AB e ABC mantém as mesmas funções: AB é responsável pelas instruções de como relacionar A com B (o *script* com a história) e ABC é responsável por nos ajudar a relacionar C com AB, nos dando a capacidade de fornecer *respostas* às preguntas realizadas.

eficácia equivalente. Contrariando as crenças dessas pessoas, o fato é que não compreendemos uma palavra de chinês e tudo o que fizemos foi manipular símbolos, uma tarefa basicamente formal, despreocupada com toda e qualquer semântica, voltada apenas para uma sintaxe.

Voltando ao programa de Schank, suposto entendedor de histórias e capaz de explicar a capacidade humana de compreender tais histórias (na visão de um defensor da IA Forte), podemos afirmar, à luz do que o AQC nos traz, que uma máquina que instancie tal programa é análoga à sala na qual nos confinamos em nossa experiência de pensamento. Diante disso, podemos extrair duas conclusões imediatas que confrontam as supostas capacidades que tal máquina teria (às quais foram citadas anteriormente, na p. 4):

1) Com relação à primeira capacidade ou reivindicação que nos foi dada, da mesma forma que, em nossa experiência de pensamento, não fomos aptos a compreender nenhuma palavra de chinês, embora obtivéssemos sucesso em nossa empreitada, um programa de computador que responde corretamente a perguntas sobre histórias não poderá jamais compreender o que são essas histórias ou qualquer outra coisa relacionada a elas, que ultrapasse a barreira do formal e se enverede pela semântica. Em suma, um computador *não entende nada*.

2) Com relação à segunda capacidade ou reivindicação (de que o computador explicaria a compreensão humana), não se pode dizer que ela se verifica, posto que nem o programa e nem a sala seriam dotados da capacidade de compreender algo, não provendo, dessa forma, subterfúgios para uma explicação de tal natureza. Segundo Searle (1980, p. 420),

> Enquanto o programa é definido em termos de operações computacionais em elementos definidos de maneira puramente formal, o que o exemplo sugere é que estes, sozinhos, não possuem nenhuma conexão interessante com a compreensão. Certamente, eles não são condições suficientes, e não foi dada a menor razão para se supor de que eles são condições necessárias ou mesmo de que realizem uma contribuição significante para a compreensão[41].

[41] "As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been

Além desses dois pontos, é preciso estar atento a mais uma reivindicação searleana: o problema não circula apenas em torno de se afirmar que diferentes máquinas poderiam possuir os mesmos *inputs* e *outputs* tendo como base diferentes princípios formais; faz-se mister a ênfase de que quaisquer que sejam os princípios formais implementados em uma máquina, nenhum deles será capaz de capacitá-la a *compreender* alguma coisa, já que um ser humano poderia seguir os mesmos princípios, se comportar como se tivesse desenvolvido a compreensão de algo e "funcionar" como se tivesse desenvolvido tal compreensão, sem que ainda assim tenha compreendido coisa alguma.

Como o leitor já pode ter percebido, o ensaio de Searle possui como *background*, dentre outros conceitos-chave, o conceito de *compreensão*, ou *entendimento* de algo. Não é possível discorrer sobre máquinas que compreendem (ou não), ou sobre o que humanos compreendem (ou não) sem que enfoquemos o próprio conceito de compreensão. O que seria, então, para Searle, a compreensão? Em primeiro lugar, a compreensão não seria um predicado binário, por assim dizer (no sentido de haver apenas estados de compreensão completa e incompreensão completa); nesse caso, o princípio do terceiro excluído seria excetuado, já que se pode identificar diversos *tipos* ou, melhor ainda, *níveis* de conhecimento/compreensão. Obviamente, os defensores da IA Forte poderiam aproveitar o ensejo desta primeira característica do conceito de compreensão e afirmar que as máquinas compreendem, mas de uma *forma diferente* ou num *nível distinto* do dos seres humanos, o que nos leva à nossa segunda característica.

O sentido que se empregaria à sentença "eu compreendo português melhor do que alemão" não é o mesmo sentido empregado à sentença "a calculadora compreende mais de matemática do que meu filho de 5 anos". Segundo Searle, é bastante comum e peculiar que atribuamos intencionalidade a objetos que foram construídos por nós para que servissem como *extensões* de nossas habilidades: máquinas de calcular, medidores de temperatura, relógios etc. Tais ferramentas podem funcionar como extensões de nossos limites ou facilitadores, mas é um equívoco acreditar que uma calculadora *compreenda* matemática, que um relógio *saiba* como contar as horas ou que um termômetro *conheça* a temperatura exata. O AQC tenciona demonstrar que um computador, não importa o programa que esteja

---

given to suppose that they are necessary conditions or even that they make a significant contribution to understanding". (SEARLE, 1980, p. 420)

implementado nele, compreende exatamente o que uma calculadora, um relógio ou um termômetro compreende, a saber, nada. Diferentemente do ser humano, que pode compreender algo de forma parcial ou incompleta, a compreensão do computador é *zero*.

2.3. O que o argumento pretende atingir

Até agora, ficou bem claro que é certo afirmar que o AQC implica a reivindicação de que uma habilidade cognitiva não pode se resumir à simples instanciação de um programa em uma máquina, ou seja, não pode ser simplificada à mera manipulação de símbolos formais. Poder-se-ia afirmar, entretanto, que o AQC atingiria apenas aquelas máquinas rodando programas de compreensão de histórias, sendo possível alterar tal realidade caso fosse instanciado um programa de *aprendizagem*, por exemplo. É preciso afirmar desde já que essa postura estaria equivocada, tendo em vista que as mesmas implicações recairiam sobre qualquer programa, seja ele de aprendizagem de línguas, de compreensão de histórias ou de qualquer outra faculdade humana. Isso porque, em quaisquer dos casos, tais habilidades estariam sendo (ou tentando ser) duplicadas a partir da simples manipulação de símbolos. É preciso enfatizar este caráter do AQC para que fique explícito o primeiro "alvo" do argumento searleano: os *programas*, todos os programas. Ainda que algum programa passe no Teste de Turing (ou seja, consiga se "fazer passar" por um ser humano, como acontece no caso do programa de Schank), isso de forma alguma garante que ele seja algo mais do que uma mera simulação de uma habilidade humana. Isso não significa, como já afirmamos mais cedo no início deste capítulo, que o foco principal do AQC sejam as máquinas; segundo Searle, é possível que máquinas possam compreender, aprender e raciocinar: o cérebro humano é um grande exemplo disso, já que o cérebro é um tipo de máquina.

Como já afirmamos no capítulo anterior (ver sessão 1.4.), e vale a pena recordar agora, o *computacionalismo* é aquela teoria que afirma que a computação é tanto necessária quanto suficiente para a cognição, pelo fato dos estados mentais serem (unicamente) estados computacionais. Dessa forma, o computacionalismo vem a ser o segundo alvo o qual o AQC pretende atingir, tendo em vista que, para

Searle, não seria possível produzir intencionalidade apenas a partir de estados computacionais.

O nosso terceiro alvo, intimamente ligado ao computacionalismo, é o *cognitivismo*, área maior em que aquele primeiro se localiza. O cognitivismo é um programa de pesquisa que atualmente predomina no campo da ciência cognitiva, sendo considerado por Searle o método contemporâneo de se fazer ciência cognitiva. De acordo com ele, (SEARLE *apud* PRESTON, 2007, p. 21),

> Pensar é processar informação, mas processar informação é apenas manipulação de símbolo. Computadores realizam manipulação de símbolo. Então, a melhor forma de estudar o pensamento (ou como os cognitivistas preferem chamar, "cognição") é estudar os programas computacionais de manipulação de símbolos, quer estejam eles em computadores ou em cérebros[42].

Não é difícil perceber o motivo pelo qual o cognitivismo se torna alvo do AQC: mais uma vez a manipulação de símbolos seria priorizada e colocada como condição necessária e *suficiente* para o desenvolvimento de habilidades que pressupõem intencionalidade.

Em quarto e quinto lugares, podemos citar como alvos do AQC o *behaviorismo* e o *funcionalismo*, por motivos bastante específicos: o behaviorismo, no nosso contexto específico, afirmaria que ao apresentar o mesmo comportamento (ou seja, *outputs*) diante dos *inputs* (ou seja, estímulos) que recebe, o computador poderia ser visto como semelhante ao cérebro humano em virtude apenas da manipulação de símbolos; no mesmo contexto, o funcionalismo defenderia que, não apenas por se comportar, mas por *funcionar* da mesma maneira que o cérebro humano (já que, se o programa fosse o correto, o quarto não apenas se *comportaria* como se tivesse entendido chinês, mas *funcionaria* – externa *e* internamente – como tal). Podemos verificar essas informações, inclusive, a partir da leitura de Preston (2007, p. 21-22):

> Versões do *funcionalismo* e da *teoria representacional da mente*, de acordo com a qual a mente está para o cérebro assim como o programa de computador está para o hardware do computador, também são abrangidos pela mira do argumento. [...] Ele pensa no funcionalismo da máquina de Turing (pelo menos) como uma

[42] "Thinking is processing information, but information processing is just symbol manipulation. Computers do symbol manipulation. So the best way to study thinking (or as they prefer to call it, 'cognition') is to study computational symbol-manipulation programs, whether they are in computers or in brains". (SEARLE *apud* PRESTON, 2007, p. 21)

> variante de uma versão na qual o próprio Putnam contribuiu um pouco para descreditar: o behaviorismo. (O behaviorismo, é claro, como se precisasse de mais alguma refutação avançada, também está na mira do AQC). O AQC deve refutar o funcionalismo porque o quarto, se os programadores fizeram tudo direito, não apenas se *comporta* como se entendesse chinês, ele *funciona* (externa *e* internamente) como se entendesse[43].

Em sexto e último lugar, temos o Teste de Turing como um dos maiores alvos do AQC. Como já apresentamos no capítulo anterior (ver sessão 1.1.), o TT seria um teste pelo qual uma máquina passaria tendo em vista atestar se esta poderia se passar, em um sentido behaviorista, por um ser humano. Segundo Searle (apud PRESTON, 2007, p. 22), muitos pesquisadores da IA Forte estariam convencidos de que o TT seria suficiente para atestar a mentalidade em dispositivos computacionais; sendo assim, se uma máquina passasse no TT demonstrando que consegue compreender chinês como um humano consegue, então poderia ser dito que ela de fato compreende. Searle aponta a falha do TT nesse aspecto afirmando que estaria se confundindo *epistemologia* com *ontologia*, de acordo com as definições do próprio autor, no que tange a tais conceitos: como seria levado em conta o aspecto comportamental, ou seja, a visão de uma terceira pessoa, apenas o caráter epistemológico estaria em voga, à medida que o crucial neste caso seria o caráter ontológico, representado pela visão em primeira pessoa, que no AQC foi bem ilustrado pelo humano dentro do quarto. O aspecto ontológico do mental pode ser bem descrito a partir da seguinte asserção de Searle (2006, p. 34):

> Porque os fenômenos mentais estão essencialmente relacionados à consciência, e porque a consciência é essencialmente subjetiva, segue-se que a ontologia do mental é essencialmente uma ontologia de primeira pessoa. Os estados mentais são sempre estados mentais de alguém. Há sempre uma "primeira pessoa", um "eu", que tem esses estados mentais. A consequência disso para a presente discussão é que o ponto de vista de primeira pessoa é primeiro. Na prática efetiva de investigação, estudaremos, é claro, outras pessoas, simplesmente porque a maior parte de nossa pesquisa não é sobre nós mesmos. Mas é importante enfatizar que o que estamos tentando atingir ao estudarmos outras pessoas é precisamente o ponto de vista de primeira pessoa. Quando estudamos *ele* ou *ela*, o

[43] "Versions of *functionalism* and the *representational theory of mind* according to which the mind is to the brain as computer program is to computer hardware also fall within the argument's sights. […] He thinks of Turing Machine functionalism (at least) as a variant on a view which Putnam himself have already done his bit to discredit: behaviourism. (Behaviourism, of course, as if it needed any further refutation, is also within the sights of the CRA). The CRA is supposed to refute functionalism because the Room, if the programmers have got things right, not only *behaves* as if it understood Chinese, it *functions* (externally *and* internally) as if it does". (PRESTON, 2007, p. 21-22)

> que estamos estudando é o *eu* que é ele ou ela. E esta não é uma questão epistêmica.

Finalmente, tendo em vista que um *Gedankenexperiment*, em nosso caso particular, funciona de modo a testar teorias da mente (voltadas à IA) tendo em vista se elas se comportam, instanciadas na mente de alguém, da mesma forma como se comportariam hipoteticamente, o AQC tenciona revelar que as teorias que enumeramos anteriormente se mostram falsas quando postas à prova de uma visão de primeira pessoa. Não obstante, alguns filósofos não se mostraram totalmente convencidos de que o AQC realmente consegue realizar esta tarefa (invalidar certas teorias); temos, então, uma lista de objeções que podem ser apresentadas ao argumento, as quais algumas foram respondidas por Searle e outras ainda estão esperando por uma réplica. Trataremos agora dessas objeções e de suas consequências.

3. AS OBJEÇÕES AO ARGUMENTO DO QUARTO CHINÊS

Compreendemos que, agora que consideramos o AQC devidamente exposto e esmiuçado, faz-se mister a apresentação e um estudo mais atencioso ao que chamamos aqui de *objeções ao argumento do quarto chinês*. Antes de tudo é preciso esclarecer que o texto contemplado em nosso trabalho (o artigo de Searle) foi estudado, debatido e criticado por muitos filósofos; por motivos pragmáticos, não será possível expor nesse momento, de forma enciclopédica, tudo o que, neste aspecto, sucedeu o AQC. Entretanto, faremos recortes pertinentes a fim de não apenas tratar este tema de maneira ampla, mas proporcionar ao leitor o instrumental necessário para analisar as diversas implicações que permeiam o argumento, bem como seus pontos fortes e fracos.

Para tanto, em primeiro lugar apresentaremos o que compreendemos por *objeções iniciais*, a saber, aquelas que são descritas originalmente por Searle em seu artigo, logo após a apresentação do AQC. As objeções iniciais serão seguidas das réplicas de Searle a tais objeções (que também estão contidas em seu texto), além das primeiras *objeções tardias*. Por objeções tardias compreendemos todas aquelas objeções que foram desenvolvidas após o lançamento do artigo searleano. Como acima dito, não apresentaremos *todas* as objeções por motivos práticos, mas as que foram selecionadas[44] podem ser divididas em dois grupos: as *objeções de cunho lógico* (que nos referimos anteriormente por “primeiras objeções tardias”), tecidas em sua totalidade por Jack Copeland, e as *objeções de conteúdo*, desenvolvidas por outros filósofos, tais como Stevan Harnad e Georges Rey. As objeções de cunho lógico serão apresentadas primeiramente (em relação às outras objeções tardias) por se referirem diretamente às objeções iniciais, para que dessa forma possamos apresentá-las de maneira metódica e proporcionando o debate (objeção-réplica-tréplica) em torno dessas primeiras críticas.

[44] Optamos por selecionar as objeções que consideramos mais relevantes em relação a nossa temática, a saber, as de maior profundidade (não superficiais) e que pudessem render um diálogo proveitoso com o texto original de Searle, por apresentarem refutações bem desenvolvidas e intrigantes.

3.1. Objeções iniciais

3.1.1. A objeção dos sistemas (The systems reply)

Temos no total seis objeções trazidas por Searle em seu artigo. A primeira delas, intitulada *a objeção dos sistemas*, versa que:

> Embora seja verdade que a pessoa que está trancada no quarto não compreende a história, ocorre que ela é meramente parte de um sistema global, e o sistema compreende a história. Essa pessoa tem uma grande tabela à sua frente na qual estão escritas as regras, tem um bloco de papel de rascunho, lápis para fazer cálculos; além disso tem um 'banco de dados' com um conjunto de símbolos em chinês. Assim sendo, a compreensão não deve ser atribuída a um simples indivíduo, mas à totalidade de um sistema do qual ele faz parte[45]. (SEARLE, 1980, p. 421).

A motivação por trás dessa primeira objeção é bastante simples: compreende-se que o problema no argumento searleano é que ele contempla apenas uma parte de um todo maior, que seria o sistema em si. De tal forma, não seria justo afirmar que um sistema não compreende algo tomando como base apenas uma parte dele; seria preciso ver o cenário maior, o sistema como um todo. Vejamos qual seria a resposta a essa objeção.

Tomando a situação através da perspectiva searleana, poderíamos até considerar tal objeção um tanto ingênua: Searle pede apenas que se internalize o sistema e suponha-se que tudo aquilo que o compreenda (além do homem que realiza cálculos), a saber, as regras, o banco de dados, os papéis e canetas, enfim, todo o resto, seja suprimido pelo indivíduo dentro da sala. (SEARLE, 1980, p. 421). Este, então, memorizaria tudo o que se haveria de memorizar, relacionando todos os símbolos "dentro de sua cabeça" e fornecendo as respostas necessárias. Sendo assim, o homem se tornaria o sistema em si, já que internalizara todos os seus elementos. A pergunta crucial neste momento é: desta feita, o indivíduo se tornou capaz de compreender chinês? A resposta a qual Searle e todos nós chegamos é a

[45] "While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has 'data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part". (SEARLE, 1980, p. 421)

mesma: não, a compreensão não se tornou possível nem mesmo através desta empreitada. Mesmo agora, o indivíduo estaria realizando apenas operações de cunho formal, dotadas de nenhum sentido ou significado, ou seja, estaria lidando apenas com a sintaxe e não com a semântica. Em suma, a réplica de Searle é que se o indivíduo internalizou o sistema inteiro e ainda assim não foi capaz de compreender chinês, o sistema jamais compreenderá chinês, pois tudo o que há no sistema também há no indivíduo agora, e se ele não consegue compreender, nenhuma parte sua também poderá; sendo o sistema uma parte sua agora, o sistema também não compreende chinês.

Talvez ainda mais importante que a sua réplica, sejam as várias implicações, apresentadas por Searle, decorrentes de uma ideologia impregnada na objeção dos sistemas. Segundo o autor, a crença de que possivelmente "[...] a conjunção de uma pessoa e pedacinhos de papel poderia compreender chinês[46]" (1980, p. 421) está ligada à ideologia da IA Forte e sugere que dentro de um só homem existem vários subsistemas: aquele que compreende chinês, o que compreende inglês, e assim em diante. O partidário de tal ideologia defenderia, então, que um subsistema se diferencia de outro, apesar de ambos abarcarem a compreensão. Para Searle, o fulcral neste momento é determinar o porquê de tal diferenciação: enquanto o subsistema que compreende inglês *conhece* hambúrgueres e o associa a restaurantes e locais afins, *compreendendo* que o nome hambúrguer é a representação linguística para um dado objeto no mundo, o subsistema que compreende chinês atua de modo diferente. Este último relaciona um rabisco de formato X a um rabisco de formato Y e por isso o indivíduo é capaz de, em vários momentos, responder corretamente a questões concernentes a tais rabiscos. Os dois subsistemas podem fornecer os mesmos *outputs* diante de *inputs* semelhantes, mas estão longe de ser a mesma coisa.

Ainda com relação às implicações decorrentes dessas crenças, a pergunta: "o que motiva a objeção dos sistemas?" é completamente pertinente. Tal pergunta pode ser lida de outra maneira: "[...] que fundamentos independentes existem para se dizer que o agente deve ter um subsistema dentro dele que literalmente compreende histórias em chinês?[47]" (SEARLE, 1980, p. 422). Searle acredita que a

[46] "[...]the conjunction of that person and bits of paper might understand Chinese". (SEARLE, 1980, p. 421)
[47] "[…] what independent grounds are there supposed to be for saying that the agent must have a

motivação para isso reside na crença de que possuímos um programa, o qual seria o tal subsistema, e tal programa teria a função de compreender histórias em chinês. Sendo um programa, poderíamos testá-lo usando o TT. Dessa forma, teríamos os mesmos *inputs* e *outputs*, mas apesar dos dois subsistemas (o que entende chinês e o que entende a língua materna) passarem no teste (ambos poderiam convencer uma segunda pessoa de que o indivíduo realmente compreende ambas as línguas), apenas um deles compreende de maneira genuína.

Persistindo na temática dos tais subsistemas, se seguirmos o mesmo raciocínio, chegaremos a conclusões absurdas: se nosso organismo pode compreender vários subsistemas, todos eles funcionando desta mesma maneira (recebendo *input* e devolvendo *outputs*), então todos os subsistemas não-cognitivos poderiam se tornar cognitivos. Tomemos como exemplo o estômago: ele processa um dado tipo de informação (a comida) recebendo *inputs* e respondendo com *outputs*, e poderíamos até chamar isso de processamento de informação, dando-lhe o crédito por instanciar, neste sentido, um programa, mas nem por isso o estômago possui compreensão (SEARLE, 1980, p. 422). A implicação, neste caso, é que, ao aceitarmos a objeção dos sistemas, poderíamos simplesmente estar concordando que órgãos como fígado, coração e estômago possuem cognição, já que funcionam da mesma forma que os outros subsistemas que já mencionamos. O que nos levaria a tal conclusão absurda seria o fato de levarmos apenas em conta o aspecto formal desse tipo de instanciação. Não se poderia sequer contestar a assimilação entre o alimento e o chinês, já que, em ambos os casos, nenhum dos dois poderia ser descrito propriamente como informação dotada de significado; seriam, portanto, nada mais que dados aos olhos dos programadores e intérpretes.

A objeção dos sistemas levanta ainda uma última discussão, na qual Searle particularmente se pronuncia de forma veemente: segundo o autor, a IA no seu sentido Forte falha em quando pretende deter o caráter de investigação cognitiva. Isto se dá por não conseguir distinguir sistemas verdadeiramente mentais daqueles não mentais. O exemplo dos subsistemas demonstra o que acabamos de afirmar: por relegar o critério de mentalidade tão somente aos olhos de um observador, a IA Forte é capaz de afirmar a existência de mentalidade com relação a um estômago ou um furacão. Searle cita McCarthy: “Podemos dizer que máquinas tão simples

---

subsystem within him that literally understands stories in Chinese?”. (SEARLE, 1980, p. 422)

como os termostatos têm crenças, e ter crenças parece ser uma característica de muitas máquinas capazes de resolver problemas[48]" (MCCARTHY *apud* SEARLE, 1980, p. 422). Tal afirmação leva Searle a acreditar que a IA Forte, como teoria da mente, está fadada ao malogro. Foquemos agora nas palavras de Searle:

> O estudo da mente começa com o fato de que seres humanos têm crenças e que termostatos, telefones e máquinas de somar não as têm. Se você concebe uma teoria que nega tal ponto, você produziu um contraexemplo e a teoria é falsa. Têm-se a impressão de que os pesquisadores da IA que escrevem esse tipo de coisa pensam que podem escapar disto porque eles realmente não levam tais coisa a sério e não pensam que alguém o fará. Proponho, pelo menos para o momento, levar estas coisa a sério. Pense por um minuto o que seria necessário para estabelecer que o pedaço de metal na parede tem, de fato, crenças – crenças com direcionalidade, conteúdo proposicional, condições de satisfação; crenças que têm a possibilidade de serem fortes ou fracas, ansiosas ou seguras, dogmáticas, racionais ou supersticiosas, fé cega ou especulações hesitantes. O termostato não é um candidato plausível a ter crenças, nem tampouco o são o estômago, o fígado, a máquina de somar ou o telefone[49]. (SEARLE, 1980, p. 423)

Como já mencionamos em capítulos anteriores, um dos aspectos da teoria searleana em torno da intencionalidade contempla a constatação de que é bastante corriqueiro acreditarmos que objetos como máquinas de calcular e termostatos possuem intencionalidade pelo fato de tais ferramentas e instrumentos constituírem uma espécie de extensão de nossa própria intencionalidade. Este erro, portanto, deve ser evitado.

---

[48] "Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance". (MCCARTHY *apud* SEARLE, 1980, p. 422)

[49] "The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this point you have produced a counterexample to the theory and the theory is false. One gets the impression that people in AI who write this sort of thing think they can get away with it because they don't really take it seriously, and they don't think anyone else will either. I propose for a moment at least, to take it seriously. Think hard for one minute about what would be necessary to establish that that hunk of metal on the wall over there had real beliefs with direction of fit, propositional content, and conditions of satisfaction; beliefs that had the possibility of being strong beliefs or weak beliefs; nervous, anxious, or secure beliefs; dogmatic, rational, or superstitious beliefs; blind faiths or hesitant cogitations; any kind of beliefs. The thermostat is not a candidate. Neither is stomach, liver adding machine, or telephone". (SEARLE, 1980, p. 423)

3.1.2. A objeção do robô (The robot reply)

A nossa segunda objeção inicial é intitulada *a objeção do robô*; em suma, é proposta uma nova experiência de pensamento distinta da originalmente ensaiada por Searle:

> Suponhamos que escrevêssemos um programa diferente daquele de Schank. Suponhamos que puséssemos um computador dentro de um robô e que esse computador não fosse apenas receber símbolos formais como *input* e produzir esses símbolos como *output*, mas que ele fosse operar o robô de tal maneira que este fizesse coisas como perceber, andar, mover-se, pregar pregos, comer, beber ou qualquer outra coisa. O robô teria uma câmara de televisão adaptada a ele – o que o capacitaria a ver -, teria braços e pernas que o capacitariam a agir e tudo isso seria controlado pelo seu cérebro-computador. Tal robô poderia ter, diferentemente do computador de Schank, compreensão genuína e outros estados mentais[50]. (SEARLE, 1980, p. 423)

O que devemos nos perguntar agora é o seguinte: o que é *necessário* para haver uma compreensão genuína e outros estados mentais? O AQC busca demonstrar principalmente o que não seria *suficiente* para atingir tais capacidades: a simples manipulação formal de símbolos. E é através da compreensão do que seria necessário e suficiente que Searle tece a sua réplica à segunda objeção. Segundo o filósofo, apesar de tal objeção levar em consideração que a sintaxe não é suficiente para a compreensão (adicionando, portanto, capacidades motoras a um robô e proporcionando-lhe interação com o mundo externo), o seu projeto mecânico não acrescentaria nada ao projeto original de Schank, já que tal robô não seria mais avançado em termos de compreensão ou intencionalidade. Dá-se isto porque o processo seria o mesmo, adicionando-se apenas mais variáveis ou, visto de outra forma, um novo elemento entre o *input* e o *output*. Tais barreiras seriam os dispositivos que proporcionariam ao robô sua percepção e movimento. Regras – e nada mais que isso – seriam seguidas em prol de fornecer informações para um

[50] "Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating drinking -- anything you like. The robot would, for example have a television camera attached to it that enabled it to 'see,' it would have arms and legs that enabled it to 'act,' and all of this would be controlled by its computer 'brain.' Such a robot would, unlike Schank's computer, have genuine understanding and other mental states". (SEARLE, 1980, p. 423)

dispositivo de vídeo ou áudio para que este, por sua vez, reagisse de maneira X ou Y, fazendo com que o robô andasse, mexesse os braços ou transmitisse imagens através de uma câmera.

A compreensão disto é facilitada pelo contraexemplo searleano:

> Suponha que em vez de um computador dentro de um robô, você me ponha dentro do quarto e me dê novamente símbolos em chinês com instruções em inglês para combinar estes símbolos com outros símbolos em chinês. Suponhamos que sem eu saber, alguns dos símbolos em chinês que chegam a mim venham de uma câmara de televisão adaptada ao robô, e que outros símbolos em chinês que estou produzindo sirvam para fazer com que o motor dentro do robô mova seus braços e pernas. É importante enfatizar que tudo que estou fazendo é manipular símbolos formais[51]. (SEARLE, 1980, p. 423)

Como dissemos, a diferença da segunda objeção é a terceirização dos processos formais aos quais já nos havíamos acostumado. Entrementes, a natureza deste processo não muda, e nada mais se faz além de manipulação de símbolos formais.

3.1.3. A objeção do simulador cerebral (The brain simulator reply)

Talvez a nossa terceira objeção seja a mais tentadora a ser defendida, por propor, como o seu título sugere, uma simulação do próprio cérebro humano. Eis a proposta:

> Suponhamos que nós projetássemos um programa que não represente a informação que temos acerca do mundo como é o caso da informação dos roteiros de Schank. O programa simula a sequência efetiva da atividade dos neurônios nas sinapses do cérebro de um falante nativo de chinês, quando este entende histórias e dá respostas a elas. A máquina recebe histórias em chinês e questões acerca delas como *input*; ela simula a estrutura formal dos cérebros dos chineses ao processar estas histórias e fornece respostas em chinês como *outputs*. Podemos até imaginar que a máquina não opera com um único programa serial, mas com

[51] "Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols". (SEARLE, 1980, p. 423)

> um conjunto de programas operando em paralelo, da mesma maneira que cérebros humanos possivelmente operam quando processam linguagem natural. Em tal caso teríamos de dizer que a máquina entenderia histórias, e se nos recusássemos a dizer isso não teríamos também que negar que falantes de chinês entendem histórias? Ao nível das sinapses, o que poderá ser diferente no programa do computador e no programa do cérebro dos chineses?[52] (SEARLE, 1980, p. 423)

Devemos agora nos interrogar: primeiramente, a compreensão e a intencionalidade podem ser reduzidas às sinapses do cérebro? Em segundo lugar, o ramo da Inteligência Artificial o qual Searle denomina IA Forte tem o mesmo escopo da neurofisiologia? Tendo em vista responder à primeira pergunta, devemos elucidar uma outra obra searleana, cujo objetivo é perscrutar o conceito de intencionalidade. Em *Intencionalidade*[53], Searle nos traz o conceito de *naturalismo biológico*, pautado na afirmação de que a consciência (ou a mentalidade) é o *resultado* do processo que ocorre no cérebro, estando intimamente ligada à matéria e dependendo da mesma, sem que para isso precise ser reduzida a esta última; a consciência, portanto, estaria para o cérebro assim como a liquidez estaria para a água, sendo ela uma propriedade emergente de um dado sistema (SEARLE, 2006, p. 162). Como, então, uma simulação cerebral que compreendesse apenas alguns de seus aspectos físicos poderia ser assimilada à mentalidade? Searle relembra o escopo original da IA Forte quando afirma que:

> Eu pensei que a ideia da IA no sentido forte é que não precisamos saber como o cérebro funciona para saber como a mente funciona. [...] Nas suposições da IA Forte, temos que a mente está para o cérebro assim como o programa está para o hardware, e portanto podemos entender a mente sem fazer neurofisiologia[54]. (SEARLE, 1980, p. 424)

---

[52] "Suppose we design a program that doesn't represent information that we have about the world, such as the information in Schank's scripts, but simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them. The machine takes in Chinese stories and questions about them as input, it simulates the formal l structure of actual Chinese brains in processing these stories, and it gives out Chinese answers as outputs. We can even imagine that the machine operates, not with a single serial program, but with a whole set of programs operating in parallel, in the manner that actual human brains presumably operate when they process natural language. Now surely in such a case we would have to say that the machine understood the stories; and if we refuse to say that, wouldn't we also have to deny that native Chinese speakers understood the stories? At the level of the synapses, what would or could be different about the program of the computer and the program of the Chinese brain?"." (SEARLE, 1980, p. 423)

[53] Cf. SEARLE, John R. *Intencionalidade*: um ensaio em filosofia da mente. Trad. Julio Fischer e Tomás Rosa Bueno. São Paulo: Martins Fontes, 2007.

[54] "I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works. […] on the assumptions of strong AI, the mind is to the brain as the program is to

Destarte, a empreitada que consiste em reproduzir parte a parte toda a estrutura cerebral, duplicando perfeitamente as causas para se atingir os mesmos efeitos, trairia o próprio objetivo da IA em seu sentido forte, já que esta se propõe a explicar o funcionamento da mente sem precisar recorrer ao cérebro, focando-se apenas no *software*.

Deixadas de lado as indagações anteriores, ainda resta um problema grave concernente ao seu conteúdo desta objeção. Searle (1980, p. 424) nos pede para comparar a simulação cerebral ao AQC, adicionando alguns elementos novos: suponhamos que o homem dentro do quarto realizasse a atividade de manipular tubulações complexas que precisassem levar a água de um canto a outro, de modo que ela entrasse por um cano e saísse por outro de forma correta. Para isso, o homem teria um programa (livro de regras) que deveria consultar para saber quais canos ligar em quais momentos e de que forma. A água, nessa adaptação, seria a informação que é recebida e que deve ser enviada (transformada, em forma de resposta) de volta após o processamento dentro da tubulação. Ainda teríamos o equivalente aos mesmos *inputs* e *outputs*: perguntas de um lado, respostas de outro. Realizando a atividade de conectar canos, abrir torneiras e manipular o conteúdo dos canos, o homem se aproximaria do processo que permeia a estrutura cerebral, referente às sinapses. Diante disso, é pertinente nos perguntarmos: como poderemos sustentar a assertiva de que tal sistema realmente compreende algo? Nem o homem e nem a tubulação compreende nada, e toda a atividade é pautada em regras que são seguidas de maneira correta devido a existência de um programa. Mais uma vez, apenas o aspecto formal foi levado em conta. Como ressaltamos no início desse tópico, à luz do pensamento searleano, para haver compreensão é necessário mais do que sinapses e neurônios, posto que a mentalidade é fruto de um processo e não pode ser reduzida à materialidade (tal pressuposto, como já indicamos anteriormente, é denominado por Searle *naturalismo biológico* [55]). Talvez o simulador cerebral esteja simulando coisas erradas: além de uma estrutura formal, ele deveria buscar simular principalmente as propriedades causais e intencionais do cérebro.

---

the hardware, and thus we can understand the mind without doing neurophysiology". (SEARLE, 1980, p. 424)

[55] Cf. SEARLE, John R. *Intencionalidade*: um ensaio em filosofia da mente. Trad. Julio Fischer e Tomás Rosa Bueno. São Paulo: Martins Fontes, 2007.

3.1.4. A objeção da combinação (The combination reply)

A objeção da combinação sugere que se todas as objeções anteriores falharam, talvez a combinação das três seja a resposta mais coerente para invalidar o Argumento do Quarto Chinês. Analisemo-la agora.

> Enquanto as três objeções anteriores podem não ser convincentes como uma refutação do contraexemplo do quarto chinês, mas se elas forem tomadas conjuntamente são convincentes e decisivas. Imagine um robô com um computador em forma de cérebro alojado em sua cavidade craniana; imagine que o computador está programado com todas as sinapses de um cérebro humano; imagine que o comportamento do robô é indistinguível do comportamento humano e agora pense nisto tudo como um sistema unificado e não apenas como um computador com *inputs* e *outputs*. Certamente em tal caso teríamos que atribuir intencionalidade ao sistema [56] . (SEARLE, 1980, p. 424)

Ao construir sua réplica, Searle parte de dois pontos principais para refutar a quarta objeção: em primeiro lugar, faz-se necessário lembrar que de acordo com a IA Forte, apenas o conjunto de programa e *inputs* e *outputs* corretos já seria condição suficiente e necessária para haver intencionalidade. Entretanto, ao assimilar o robô descrito a um humano com intencionalidade genuína, esses processos formais que mencionamos não seriam levados em consideração, apenas a semelhança motora do artefato. “Se o robô parece conosco, ele deve ter intencionalidade e pensar como nós”, assumiríamos. E seria absolutamente natural pensar dessa maneira enquanto não conhecêssemos a verdadeira natureza do robô e a maneira através da qual ele é operado. Entretanto, se todos conhecessem a engenhoca interna que, através de regras formais, aciona um mecanismo ou outro, dando movimento e outras reações ao robô, ainda acreditariam que o mesmo possui intencionalidade no mesmo sentido em que nós a possuímos? A quarta objeção, nesse sentido, comete o mesmo erro que a segunda.

---

[56] “While each of the previous three replies might not be completely convincing by itself as a refutation of the Chinese room counterexample, if you take all three together they are collectively much more convincing and even decisive. Imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behavior of the robot is indistinguishable from human behavior, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system”. (SEARLE, 1980, p. 424)

Em segundo lugar, se imaginássemos (assim como fizemos anteriormente com as demais objeções) que em vez do computador, um homem estivesse dentro do robô, realizando toda e qualquer atividade que um computador poderia realizar, assim como supomos nas primeiras três objeções, ainda assim não poderíamos encontrar nem compreensão e nem intencionalidade – nem no homem, com relação à atividade que desempenhava, e nem no sistema como um todo – posto que nada mais do que manipulação formal seria executada. Em outras palavras,

> A manipulação de símbolos formais continua, o *input* e o *output* são combinados corretamente, mas o único *locus* de intencionalidade é o homem, e ele não sabe nada dos estados intencionais relevantes; por exemplo, ele não vê o que chega aos olhos do robô, ele não tem a intenção de mover o braço do robô, ele não compreende as observações que são feitas pelo robô ou que lhe são feitas. Nem tampouco, pelas razões colocadas acima, o sistema do qual o homem e o robô são parte, compreende alguma coisa[57]. (SEARLE, 1980, p. 425)

3.1.5. A objeção das outras mentes (The other minds reply)

Esta quinta objeção foi respondida de maneira bastante breve por Searle. Em suma, a objeção afirma que a única razão pela qual atribuímos faculdades mentais às pessoas ao nosso redor, como intencionalidade, crenças, emoções e compreensão, é devido ao seu comportamento. Diante disso, se um computador se comportar de maneira tão convincente quanto um ser humano, por que não deveríamos atribuir-lhe o mesmo que atribuímos ao humano?

Searle considera que o fazer do cientista cognitivo comporta um pressuposto básico, que é a realidade e a possibilidade de se conhecer estados cognitivos, ou seja, o mental. Não faz sentido versar sobre o tema como se tais estados fossem incógnitas e jamais pudéssemos conhecer nada a respeito deles, e por isso apenas nos mantermos desenvolvendo conjecturas sobre algo incerto apenas por hábito. Segundo o autor, “Em ‘ciências cognitivas’ pressupõe-se a realidade e a possibilidade de se conhecer o mental, da mesma maneira que em ciências físicas

[57] “The formal symbol manipulations go on, the input and output are correctly matched, but the only real locus of intentionality is the man, and he doesn't know any of the relevant intentional states; he doesn't, for example, see what comes into the robot's eyes, he doesn't intend to move the robot's arm, and he doesn't understand any of the remarks made to or by the robot. Nor, for the reasons stated earlier, does the system of which man and robot are a part”. (SEARLE, 1980, p. 425)

tem-se de pressupor a realidade e a capacidade de se conhecer objetos físicos[58]" (SEARLE, 1980, p. 425).

3.1.6. A objeção das várias casas (The many mansions reply)

A objeção das várias casas é focada numa aposta no futuro e na crença do potencial da IA Forte: acredita-se que a IA ainda não produziu um computador que duplicasse o mental porque a tecnologia não é avançada o suficiente nesta dada época, mas que isso será possível um dia. Ainda mais que isso, essa objeção supõe que a IA, futuramente, não implantará uma intencionalidade genuína apenas em computadores digitais, mas em diversos objetos que poderão exibir mentalidade: "[...] eventualmente seremos capazes de construir dispositivos que exibirão esses processos causais e isto será também inteligência artificial[59]" (SEARLE, 1980, p. 425).

Searle responde brevemente esta objeção, afirmando, em suma: a IA Forte não é simplesmente qualquer coisa que explique a cognição. A proposta original da IA Forte é que processos mentais são processos computacionais sobre elementos formais definidos, portanto não é possível generalizar a proposta da IA ao ponto de aplicá-la a qualquer coisa; dessa forma tal objeção trivializa a proposta da IA no sentido forte.

Para finalizar suas réplicas, Searle repassa algumas noções gerais contidas em seu artigo, deixando claro pontos cruciais de sua tese: em primeiro lugar, não é a instanciação de um programa que deve ser condenada (até porque somos instanciações de programas de computador [60]) mas a desconsideração da

[58] "In 'cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects". (SEARLE, 1980, p. 425)

[59] "[...] eventually we will be able to build devices that have these causal processes, and that will be artificial intelligence". (SEARLE, 1980, p. 425)

[60] Searle afirma isto em seu artigo *Minds, brains and programs*, e consideramos necessário transcrever na íntegra a afirmação searleana: "Em princípio, não vejo razão pela qual nós não pudéssemos dar a uma máquina a capacidade de compreender inglês ou chinês, desde que, em um sentido importante, nossos corpos juntamente a nossos cérebros são, precisamente, tais máquinas. Mas eu vejo muitos argumentos fortes para dizer que nós não poderíamos dar tal coisa a uma máquina onde a operação da mesma é definida unicamente em termos de processos computacionais sobre elementos formais definidos; isto é, onde a operação da máquina é definida como uma instanciação de um programa computacional. Não é porque eu sou a instanciação de um programa de computador que eu estou apto a compreender inglês e ter outras formas de intencionalidade (eu sou, eu suponho, a instanciação de qualquer número de programas computacionais), mas até onde sabemos, isto acontece porque eu sou um tipo de organismo com uma certa estrutura biológica (por

necessidade de existência de um organismo com características tais que o possibilite a desenvolver uma intencionalidade. Esse organismo é uma estrutura que pode, como é o nosso caso, ser biológica, mas que não deve ser resumida a aspectos meramente formais. Não se deve levar em consideração apenas a organização formal do cérebro, a saber, o que concerne à matéria e modo através do qual essa matéria se organiza, mas deve-se visar também – e principalmente – as *propriedades efetivas* [61] de tal conjuntura. A formalidade que mencionamos é tida, para Searle, como uma mera sombra da cognição, que frequentemente é confundida com esta última.

Além disso, outros pontos centrais são postos em: respondendo à pergunta inicialmente suscitada por Turing, Searle afirma que sim, uma máquina pode sim pensar: o maior exemplo disso somos nós, seres humanos. Poderíamos, também, construir máquinas artificiais que pudessem pensar, contanto que essas máquinas tivessem estruturas semelhantes a nossa própria. Contanto que as causas fossem duplicadas de maneira exata, o efeito também seria. O único empecilho seria, no entanto, se quiséssemos duplicar a mentalidade partindo apenas da instanciação de um dado programa e encarássemos tais condições como necessárias se suficientes para a compreensão ou até mesmo intencionalidade. Segundo o autor, "[...] a manipulação de símbolos formais por si só não tem intencionalidade: eles não têm significado, eles nem mesmo são manipulações de símbolos, uma vez que esses símbolos não simbolizam nada[62]" (SEARLE, 1980*,* p. 427).

Sendo assim, o AQC é eficiente por adicionar, num ambiente sem intencionalidade, um sujeito detentor da mesma (a saber, um homem), e submetê-lo a um conjunto de experiências (a manipulação dos símbolos); após o término de tal atividade, o homem deixará o ambiente intacto, sem nenhuma compreensão a mais

---

exemplo, química e física), e essa estrutura, sob certas condições, é causalmente capaz de produzir percepção, ação, compreensão, aprendizado e outros fenômenos intencionais. (SEARLE, 1980, p. 426).

[61] Searle não define, explicitamente, o que seriam tais "propriedades efetivas". Supomos aqui que tais propriedades englobem, por exemplo, a intencionalidade, *qualia* e a capacidade de gerar consciência. Searle menciona tais propriedades efetivas em algumas passagens de seu texto, citaremos uma delas: "A descoberta mais surpreendente que eu fiz ao discutir estes problemas é que muitos pesquisadores da IA estão chocados com a minha ideia de que fenômenos mentais humanos podem ser dependentes das efetivas propriedades físico-químicas dos cérebros humanos" (SEARLE, 1980, p. 429)

[62] "[...] the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything". (SEARLE, 1980, p. 427)

sobre coisa alguma.

O programa jamais poderá ser confundido com a mente em si, já que é possível instanciar o mesmo programa em dispositivos absurdos como computadores construídos a partir de rolos de papel higiênico e pedras, canos d'água ou mesmo um conjunto de cata-ventos. Não é possível extrair intencionalidade a partir disso, apesar de ser possível instanciar um tipo simples de programa dessa maneira. Mais uma vez, apenas algo que tenha os mesmos poderes causais do cérebro seria capaz de tal feito. Além disso, diferente dos programas, os estados intencionais não são formais, são antes de tudo definidos a partir de seu conteúdo em vez de sua forma. Podemos adicionar a isso o fato de que à medida que os estados mentais são produtos de operações cerebrais, o mesmo não pode se dizer dos programas em relação ao computador.

Finalmente, Searle destaca vários mal-entendidos que levam os leitores e pesquisadores a considerar de maneira equivocada as questões que perscrutam o mental: frequentemente se confunde simulação com duplicação, principalmente no que concerne a qualquer atividade ou evento mental (não é comum acreditar que a simulação computacional de uma tempestade irá, realmente, criar uma tempestade, mas se acredita que a simulação do mental possa resultar na mentalidade de fato); acredita-se que o processamento de informação computacional é o mesmo que o processamento mental, sendo que este detém características que remetem significado, valor e intencionalidade, ao passo que aquele detém tão somente uma simbologia vazia (de significado). Some-se a isto uma carga de behaviorismo ou operacionalismo[63] e dualismo impregnada nas considerações da IA Forte, a partir, por exemplo, do TT, que sugerem que comportar-se e ser são dois lados de uma realidade causal e que o programa está apartado da máquina da mesma forma que a mente está apartada do cérebro (ou da matéria), podendo-se isolar um do outro, analisando-o separadamente, sem jamais levar em consideração a sinergia proveniente de um conjunto ou, melhor colocando, de uma *unidade*. O arremate final das considerações searleanas consiste em afirmar que a IA Forte tem pouco a dizer sobre o pensamento por, igualmente, ter bastante pouco a dizer sobre as máquinas

[63] *Operacionalismo* é uma outra nomenclatura utilizada por Searle (1980, p. 428-429) para designar o behaviorismo impregnado no texto de Turing, referente às máquinas.

(SEARLE, 1980, p. 429) e, consequentemente, dessa unidade à qual nos referimos: a IA forte trata tão somente de programas, e programas não são máquinas.

3.2. Objeções tardias

3.2.1. Objeções de cunho lógico

Nesta sessão apresentaremos alguns pontos traçados por Jack Copeland em seu artigo *The chinese room from a logical point of view* (2007). Copeland, assim como Searle, considera fracas as réplicas ao seu argumento original (o AQC), e busca analisá-lo logicamente (deixando um pouco de lado a semântica e preocupando-se com o conteúdo formal da argumentação). Segundo Copeland, existem quatro versões do AQC, às quais ele nomeia: versão baunilha, a versão externa, a versão do simulador e a versão do ginásio. Entrementes, apenas as duas primeiras versões estão presentes no artigo *Minds, brains and programs* de 1980, e dizem respeito unicamente à crítica a manipulação de símbolos formais que caracteriza a proposta da IA Forte; as duas últimas versões têm como base o artigo searleano *Is the brain's mind a computer program?* (2007), de 1990, marcado por uma forte crítica ao conexionismo. Além dessas quatro versões, há também uma observação relativa à chamada Tese de Church-Turing, mais especificamente à maneira equivocada que Searle a interpreta. Copeland de início já deixa claro o seu parecer sobre o AQC, considerando-o insatisfatório em todas as suas versões, além de sequer conseguir explicar de que modo o "puramente formal" ou "sintático" seria incapaz de ser constitutivo e/ou suficiente para a mente (ou o mental). Analisemos agora sobre o que versa o texto de Copeland.

3.2.1.1. A versão baunilha (The vanilla argument)

Esta versão é relativa à resposta de Searle referente a uma das objeções ao AQC, a saber, a primeira delas, a objeção dos sistemas. Copeland cita Searle:

> |Clerk| [64] não entende uma palavra das histórias chinesas. |Clerk| possui *inputs* e *outputs* que são indistinguíveis daqueles dos falantes

[64] Clerk seria o funcionário humano dentro do quarto manipulando os símbolos.

> nativos chineses, e |Clerk| pode possuir qualquer programa formal que você queira, mas |Clerk| ainda não entenderá nada. O computador de Schank, pelos mesmos motivos, não entende nada de história alguma... Quaisquer que sejam os princípios puramente formais que você ponha num computador, não serão suficientes para o entendimento, já que o humano estará apto a seguir os princípios formais, sem que haja compreensão...[65] (SEARLE *apud* COPELAND, 2007, p. 110)

Searle chega a essa conclusão por buscar, através da experiência de pensamento do AQC, uma forma de demonstrar que um funcionário humano poderia realizar os mesmos passos que o programa de Schank realizaria, e ainda assim não compreender nada de chinês. A tática de Searle, nesse momento, é tirar o foco da visão de uma terceira pessoa (que observa o programa funcionando normalmente, como se compreendesse tudo), e recoloca-lo sob a ótica de uma primeira pessoa. Esse é, precisamente, um dos objetivos da experiência de pensamento. Para justificar essa atitude (de utilizar o ponto de vista de uma primeira pessoa), Searle aponta:

> O caráter de terceira pessoa da epistemologia não nos deve cegar para o fato de que a ontologia efetiva dos estados mentais é uma ontologia de primeira pessoa. O modo como o ponto de vista de terceira pessoa é aplicado na prática torna difícil para nós perceber a diferença entre algo que realmente tem uma mente, como um ser humano, e algo que se comporta *como* se tivesse uma mente, como um computador. [...] Crenças, desejos etc., são sempre crenças e desejos *de alguém*. (2006, p. 28-29)

Copeland identifica, nessa passagem específica do texto de Searle, de 1980, alguns equívocos de cunho lógico, afirmando que o argumento não é *logicamente válido*. Segundo Copeland,

> A proposição de que a manipulação de símbolos formais carregada por Clerk não o capacita a compreender histórias chinesas não significa que implica a proposição, bastante diferente, de que a manipulação formal de símbolos carregada por Clerk não capacita a sala a entender as histórias chinesas. Pode-se, assim, afirmar que a

---

[65] "|Clerk| do[es] not understand a word of the Chinese stories. |Clerk| ha[s] inputs and outputs that are indistinguishable from the native Chinese speaker, and |Clerk| can have any formal program you like, but |Clerk| still understand[s] nothing. Shank's computer for the same reasons understands nothing of any sotries. . . . [W]hatever purely formal principles you put into the computer will not be suficiente for understanding, since a human will be able to follow the formal principles without understanding [...]". (SEARLE *apud* COPELAND, 2007, p. 110)

> declaração ‘a organização da qual Clerk é uma parte não possui bens tributáveis no Japão’ segue logicamente da declaração ‘Clerk não possui bens tributáveis no Japão’[66]. (2007, p. 110)

Percebemos, então, a falha da qual Copeland se referia: tomando o argumento de maneira estritamente formal, não seria razoável deixar de atribuir uma característica ao todo unicamente em virtude da ausência de tal característica de uma de suas partes. Refutando, assim, uma versão do AQC, localizada na réplica searleana à objeção dos sistemas, Copeland elabora uma objeção que nomeia de *objeção lógica*, e deixa clara a sua diferença em relação à objeção dos sistemas.

A objeção dos sistemas reivindica que “Enquanto é verdade que o indivíduo que é trancado na sala não compreende a história, o fato é que ele é meramente parte de todo um sistema e o sistema entende a história[67]” (SEARLE, 1980, p. 419). Podemos perceber aqui que quando migramos até o campo da semântica, tem-se um absurdo: a pessoa não entende, mas o sistema entende, tornando fácil a empreitada para aqueles que buscam refutar tal objeção. Por isso, segundo Copeland, a objeção dos sistemas é sem valor. Diferentemente disso, a objeção lógica não se preocupa com o lado semântico, apenas com o aspecto formal, por isso não se importa com a conclusão de que uma sala possa compreender chinês.

3.2.1.2. A versão externa (The outdoor version)

Ao apresentar o que chama de “a versão externa” do AQC, Copeland informa como é possível validar um argumento logicamente invalido apenas adicionando algumas premissas ao argumento, em último caso, uma dessas premissas poderia ser a conclusão. Na objeção dos sistemas, Searle defende que se uma característica é atribuída ao todo, deve ser feito o mesmo com as partes (ou a parte):

---

[66] “The proposition that the formal symbol manipulation carried out by Clerk does not enable Clerk to understand the Chinese story by no means entails the quite different proposition that the formal symbol manipulation carried out by Clerk does not enable the Room to understand the Chinese story. One might as well claim that the statement The organization of which Clerk is a part has no taxable assets in Japan’ follows logically from the statement ‘Clerk has no taxable assets in Japan’”. (COPELAND, 2007, p. 110)

[67] “While is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system and the system does understand the story”. (SEARLE, 1980, p. 419)

> Da mesma forma, ele não entende nada de chinês, e a fortiori nem o sistema entende, porque não há nada no sistema que não haja nele. Se ele não entende, então não há nenhuma maneira do sistema poder entender, porque o sistema é apenas uma parte dele[68]. (SEARLE, 1980, p. 419).

Separando as premissas e a conclusão deste argumento, temos que:

1) O sistema é parte de Clerk.
2) Se Clerk (chamemos Clerk de X) não compreende a história em chinês (chamemos "compreender a história" de Φ; por consequência, "não compreender a história", que nega o que anteriormente chamamos de Φ, chamaremos de ¬ Φ), então nenhuma parte de Clerk (x) compreende a história em chinês (Φ).
3) A manipulação de símbolos formais realizada por Clerk não o capacita a compreender a história em chinês.

Dessa forma,

4) A manipulação de símbolos formais realizada por Clerk não capacita o sistema a entender a história em chinês.

Este é, precisamente, o argumento intitulado, por Copeland, de "a versão externa", sendo considerado por ele como logicamente válido. Entretanto, na realização de uma análise lógica mais profunda dos termos que compõem as premissas, é possível encontrar vários problemas na argumentação searleana. Com relação à premissa 2, a qual Copeland nomeia de "parte-do-princípio" (*Part-of-principle*), é notado que mesmo Searle agindo como se fosse[69], essa premissa não é auto-evidente. É a partir dessa característica da premissa que será desenvolvida uma argumentação contra essa versão do argumento searleano.

---

[68] "All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way the system could understand, because the system is just a part of him". (SEARLE, 1980, p. 419)

[69] Podemos assumir que Searle toma essa premissa como auto-evidente tendo em vista a ausência de explicação com relação ao fato de nenhuma parte de Clerk compreender chinês porque Clerk, como um todo, não compreende chinês. Searle simplesmente assume que essa premissa é verdadeira sem lançar mão de nenhuma prova anterior.

A premissa 2 versa que, basicamente, se o todo não compreende alguma coisa, consequentemente nenhuma de suas partes compreenderá. A fim de demonstrar que essa concepção estaria equivocada, Copeland sugere que imaginemos que na mente de Clerk pode haver um módulo com finalidade específica de produzir soluções para certas equações tensoriais, mesmo que Clerk não faça a mínima ideia de como resolver questões desse tipo ou sequer saiba o que é uma equação tensorial. Em resumo, neste caso, admitimos que o fato de Clerk não fazer a mínima ideia de como resolver esse tipo de questão não o exime de possuir uma parte de seu cérebro capaz de resolver questões assim. Copeland afirma que Searle deixa esse ponto e o *Part-Of-Principle* totalmente sem suporte.

Copeland informa também que, de certo modo, pode-se verificar uma tentativa de Searle no sentido de dar suporte a essas questões mal ou não respondidas: o filósofo afirma que "não há nenhuma maneira de Clerk poder vir a entender chinês na situação descrita, à medida de forma alguma de Clerk pode aprender o significado de quaisquer dos símbolos[70]" (SEARLE *apud* COPELAND, 2007, p. 112-113). Com relação a isso, Copeland aparenta alegar que Searle comete uma petição de princípio, já que a afirmação de que não é possível conhecer o significado de algo através de símbolos é a motivação principal do argumento do quarto chinês.

3.2.1.3. A versão do simulador (The simulator version)

A versão do simulador é considerada por Copeland falaciosa. Podemos ilustrar tal versão a partir da seguinte citação:

> Computacionalmente, sistemas seriais e paralelos são equivalentes: qualquer cálculo que pode ser feito em paralelo pode ser feito em série. Se o homem na sala é computacionalmente equivalente a ambos, então ele não entende chinês apenas realizando cálculos, e nem eles[71]. (SEARLE *apud* COPELAND, 2007, p. 113)

[70] "There is no way |Clerk| could come to understand Chinese in the [situation] as described, since there is no way that |Clerk| can learn the meanings of any of the symbols". (SEARLE *apud* COPELAND, 2007, p. 112-113)

[71] "Computationally, serial and parallel systems are equivalent: any computation that can be done in parallel can be done in serial. If the man in the Chinese room is computationally equivalent to both, then if he does not understand Chinese solely by virtue of doing the computations, neither do they". (SEARLE *apud* COPELAND, 2007, p. 113)

Em outras palavras, Searle afirma que um sistema em série e um sistema operando paralelamente a outro são a mesma coisa, ou seja, qualquer cálculo que for feito de forma paralela, pode ser feito em série também. O funcionário humano, dentro do quarto chinês, é computacionalmente equivalente tanto a um sistema em série quanto a um sistema em paralelo. A partir de então, Searle argumenta: o homem é equivalente aos sistemas; o homem não compreende chinês; se o homem não compreende chinês, tais sistemas também não compreendem.

A partir desse raciocínio de Searle, Copeland constrói a sua crítica, trabalhando com o conceito de Máquina de Turing. Ele afirma que a crença de que várias máquinas de Turing operando em série podem simular uma Máquina de Turing Universal (*Universal Turing Machine*) é *falsa*. Isso porque máquinas de Turing universais trabalhando de maneira assíncrona são irredutivelmente paralelas, à medida que os seus processadores individuais não podem ser intercalados para formar uma sequência de ações que normalmente são desenvolvidas por uma única máquina de Turing universal. Ou seja: sistemas seriais e paralelos *não* são equivalentes. Isso se aplicaria apenas a casos em que as máquinas estivessem operando em sincronia. Pode-se compreender melhor a objeção de Copeland desmembrando o argumento de Searle, onde temos, segundo aquele, uma situação em que existe um sistema serial chamado de *S* simulando um sistema paralelo, chamado de *N*. Vejamos:

1) Dado o programa apropriado, Clerk (juntamente com suas canetas, papéis, borrachas e livros de regras) é ‘computacionalmente equivalente’ a *N*.
2) Clerk não entende chinês (se for unicamente realizar cálculos).

Então,

3) *N* não entende chinês (se for unicamente realizar cálculos).

Searle procura nos convencer de que como S simula N, o que é associado a um, deve ser associado a outro, mas isso é falso, pelos motivos acima mencionados. Copeland usa um dos argumentos de Searle contra ele mesmo: em seu texto, Searle afirma que as simulações não podem ser confundidas com

duplicações; para provar que esse raciocínio é falacioso, Searle aponta a seguinte cadeia argumentativa:

4) *x* é uma simulação de *y*; *y* tem a propriedade Φ, então, *x* tem a propriedade Φ.

Segundo Searle, dar suporte a este tipo de raciocínio é cair em erro, já que simulações são distintas de duplicações; caso contrário, uma simulação computacional de uma tempestade deixaria pessoas molhadas, e isso seria absurdo. Copeland chama essa falácia apontada por Searle de "*falácia da simulação*", e em seguida demonstra que o próprio autor comete tal erro, pois se formos substituir os termos dessa falácia por processos computacionais, teríamos que

5) *x* é uma simulação de *y*; não é o caso que *x* tenha a propriedade Φ, portanto não é o caso que *y* tenha a propriedade Φ.

Ou, em outras palavras,

6) *S* é uma simulação de *N*; não é o caso que *S* entende chinês, portanto não é o caso que *N* entende chinês.

Sendo assim, não é possível afirmar, sem que se recaia em equívoco, que uma manobra conexionista seria inválida apenas por afirmar que a simulação de Clerk poderia compreender chinês ao passo que Clerk não, já que simulações não devem ser confundidas com duplicações.

3.2.1.4. A versão do ginásio (The chinese gym)

Na versão do ginásio, Copeland acusará de falacioso o argumento do ginásio dado por Searle, em sua argumentação. Tal experiência de pensamento consiste apenas em trocar a experiência de pensamento do quarto chinês pelo que Searle

chama de “ginásio chinês”, onde há várias pessoas, todas elas falantes do inglês e totais ignorantes com relação à língua chinesa. Essas várias pessoas se encontram nesse ginásio, formando um corredor, e ficam passando por elas vários símbolos, de cores distintas: os símbolos verdes simbolizariam os *inputs* e os símbolos vermelhos simbolizariam os *outputs*. As pessoas possuiriam também uma lista, dando detalhes sobre quais símbolos eles deveriam passar à frente, para a próxima pessoa da fila, e quais deveriam ser descartados; além disso, haveria também um treinador que ficaria gritando para essas pessoas algumas instruções. Esse quadro simularia o funcionamento de um sistema, onde cada indivíduo seria um elemento constitutivo de um sistema maior, e trabalhando em equipe, formavam uma rede, se conectando entre si. Tal experiência de pensamento busca simular um sistema conexionista. O que muda nessa segunda experiência é que em vez de um homem sozinho dentro de um quarto, temos um conjunto de várias pessoas trabalhando juntas em um ginásio. O restante da experiência permanece a mesma: são feitas perguntas em chinês para que as pessoas (no caso, o sistema como um todo) respondam em chinês. Mais uma vez, o resultado também será o mesmo: *parecerá* que as perguntas foram compreendidas e respostas em chinês foram liberadas (*outputs*), dando a entender para o interrogador, que desconhece o funcionamento interno do sistema, que as partes constitutivas daquele sistema não são totalmente ignorantes com relação à língua chinesa. Vejamos de forma mais detalhada como Searle apresenta essa nova experiência de pensamento:

> Imagine que em vez de um quarto chinês, eu tenha um ginásio chinês: um corredor contendo vários falantes do inglês. Esses homens pode realizar as mesmas operações que os nodos e sinapses numa arquitetura conexionista... e o resultado poderia ser o mesmo do que o de um homem manipulando símbolos de acordo com um livro de regras. Ninguém no ginásio fala uma palavra de chinês... ainda assim, com os ajustes apropriados, o sistema poderia dar respostas corretas a questionamentos em chinês[72]. (SEARLE *apud* COPELAND, 2007, p. 116)

Copeland identifica uma falácia muito comum, imersa no discurso Searleano:

[72] “Imagine that instead of a Chinese room, I have a Chinese gym: a hall containing many monolingual English-speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture […] and the outcome would be the same as having one man manipulate symbols according to a rule book. No one in the gym speaks a word of Chinese […] Yet with appropriate adjustments, the system could give the correct answers to Chinese questions”. (SEARLE *apud* COPELAND, 2007, p. 116)

a de confundir a parte com o todo. Apesar de cada membro do sistema não compreender nenhuma palavra de chinês, essa premissa não garante que o próprio sistema, como um todo, não entenda também. A versão do ginásio consiste em duas inferências:

1) Nenhum jogador individual compreende chinês (apenas realizando cálculos).

Portanto,

2) A simulação como um todo – chamemos isto de G – não compreende chinês (apenas realizando cálculos).

Sendo assim,

3) A rede sendo simulada, *N*, não entende chinês (apenas realizando cálculos).

Como 1. não é vinculada a 2., esta primeira sentença torna-se não-sólida, ou seja, mal fundamentada. Dizemos isto porque, como foi dito anteriormente, mesmo que uma parte do todo não possua uma característica x, isso não garante que o todo não possua. Não é possível concluir 2 a partir de 1 por este motivo. Além disso, dificilmente a segunda sentença sozinha seria suficiente para os objetivos de Searle, por não ser auto-evidente. Não se pode, sem que se demonstre através de argumentos, simplesmente afirmar que uma simulação não compreende algo. É possível afirmar isto, como *conclusão*, após desencadear-se um conjunto de premissas que culminam naquela ideia final, a saber, a de que G não compreende chinês apenas realizando cálculos. Assumir como verdadeira tal asserção no meio de uma cadeia argumentativa, como se tal sentença fosse axiomática, segundo Copeland, é um erro.

Copeland também acusa Searle de se recusar a estudar a fundo as ciências cognitivas, desenvolvendo um estudo mais aprofundado com relação às ciências

físicas e biológicas. Ele afirma isso por criticar a alegação de Searle de que seria absurdo afirmar que uma mente pode ser duplicada a partir de um sistema computacional, já que, de acordo com Searle, computadores podem ser construídos até mesmo a partir de canos de água, pombos ou qualquer outra coisa, então a mente também seria duplicada a partir disso – tal ideia, segundo Searle, é absurda. Copeland, por outro lado, afirma que essa teoria (de que a réplica de um cérebro, a partir de um computador digital, possuindo um dado sistema capaz de replicar tudo que um cérebro orgânico faça, pode *realmente* ser feito de pombos) é falsa – apenas uma teoria absurda poderia implicar isto. Analogamente, da mesma forma que o conexionismo não pode implicar que um time de ginastas possa *realmente* simular o cérebro, um computador feito de pombos não pode *realmente* simulá-lo. Todavia, Copeland não provê nenhuma referência que possa sedimentar as suas afirmações. Segundo o autor,

> Quando um dispositivo computacional, *x*, se diz simular outro, *y*, frequentemente é o caso de que apenas a relação de *input-output* seja duplicada: com relação a outras coisas, os processos realizados por *x* e *y* podem diferir consideravelmente[73]. (2007, p. 118)

No caso de Searle, *G* aparenta imitar *N* muito mais do que neste quesito mencionado acima (entrada e saída de informações). *G* e *N* estão realizando o mesmo processo computacional, e podemos expressar isso dizendo que *G* é isomórfico a *N*, o que implica que a comparação de Searle (sobre pombos e canos de água) é, no mínimo, exagerada. Não se trata apenas de entrada e saída de informações, mas da réplica de um mesmo *processo*. Além disso, ainda podemos afirmar que a inferência de 1 a 3 são trivialmente válidas, por termos que "se um dado processo computacional é insuficiente para dar-se a compreensão, então ele é insuficiente para a compreensão![74]" (COPELAND, 2007, p. 118). Teríamos, então, a repetição da mesma ideia sendo tratada como se fosse algo diferente, como se de uma premissa X extraíssemos uma conclusão Y, mas na verdade 1 e 3 não se completam e nem poderiam se completar, pois tratam da mesma coisa.

---

[73] "When one computing device, *x*, is said to simulate another, *y*, it is often the case that only the input-output relationship is duplicated: in other respects, the processes carried out by *x* and *y* may differ considerably". (COPELAND, 2007, p. 118)

[74] "[…] if a given computational process is insufficient for understanding then it is insufficient for understanding!". (COPELAND, 2007, p. 118)

Finalizamos com a versão do ginásio a nossa exposição das objeções de cunho lógico apresentadas ao AQC. Como deixamos claro no início dessa seção, tais objeções se preocupam mais com a *forma* do argumento, e consideramos que tais críticas cumpriram seu papel quanto à análise da forma. Nos resta compreender de que modo o AQC pode ser objetado quanto ao seu conteúdo.

3.2.2. Objeções de conteúdo

3.2.2.1. Harnad e o computacionalismo

Talvez a mais forte objeção de conteúdo que encontraremos seja aquela relacionada aos trabalhos de Alan Turing: Searle muitas vezes é acusado de ter compreendido mal a real proposta de Turing ao inaugurar o conceito de máquina de Turing e do próprio TT. No que concerne a essa crítica, podemos trazer à tona um dos artigos de Stevan Harnad, intitulado *Minds, Machines and Searle 2: what's right and wrong about the chinese room argument*[75].

Daremos enfoque ao conceito de IA Forte e Computacionalismo, fortemente criticados por Searle. Harnad resume, em três premissas, qual seria a proposta da versão forte da Inteligência Artificial, segundo Searle:

1. A mente é um programa de computador;
2. O cérebro é irrelevante (para os estados mentais);
3. O TT é decisivo

O objetivo do AQC seria, segundo Harnad, refutar esses três princípios. Ainda segundo o autor, o máximo que se poderia fazer seria refutar a conjunção de tais princípios, o modo pelo qual se conectam, já que, tomados isoladamente, alguns desses princípios ainda poderiam ser verdade mesmo que o AQC seja válido.

Segundo o autor, "Computacionalismo é a teoria de que a cognição é computação, que estados mentais são apenas estados computacionais [76] " (HARNAD, 2007, p. 297). Pressupondo que 1 e 3 são princípios computacionalistas,

[75] Cf. HARNAD, Stevan. Minds, Machines and Searle 2: what's right and wrong about the chinese room argument. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

[76] "Computationalism is the theory that cognition is computation, that mental states are just computational states". (HARNAD, 2007, p. 247)

Harnad busca reconstruí-los a fim de apontar que Searle os compreendeu de forma equivocada. A definição de computacionalismo está intimamente ligada à premissa 1; podemos compreendê-la também da seguinte forma: estados mentais são implementações do programa computacional adequado, sendo, dessa forma, estados computacionais. Em contrapartida, disto não pode ser inferido que símbolos escritos num pedaço de papel são estados mentais: aí se encontra o erro de Searle, segundo Harnad. Os estados mentais, considerados estados computacionais (por um computacionalista), só poderiam ser gerados a partir da implementação do *código certo* no *hardware certo*, que deveria ser um sistema completamente dinâmico. O computacionalista não afirma sequer que esse "código perfeito" ou "ideal" exista, apenas que, para dar cabo de tal tarefa, os requisitos mínimos seriam estes (a saber, possuir o código e o hardware ideais).

Sobre a premissa 2, um equívoco maior ainda teria sido cometido por Searle. Em primeiro lugar, seria absurdo afirmar que o cérebro seria irrelevante para os estados mentais, já que estes acontecem na estrutura do cérebro. O que Searle gostaria de dizer, talvez, seria que tais estados mentais (que são computacionais, de acordo com a teoria computacionalista) são *independentes de implementação*. Aqui poder-se-ia entrar em conflito, à primeira vista, com a premissa 1, que afirma que estados mentais seriam a implementação do software correto no hardware correto, mas essa confusão se daria apenas nos primeiros momentos da compreensão. Refaçamos a premissa 2 da seguinte forma: estados mentais são independentes de implementação, ou seja, o software é independente do hardware.

A que nível se daria tal independência? – é algo que deveríamos nos perguntar agora. Os programas computacionais devem ser independentes de hardware no sentido de serem estruturados de modo que possam ser executados em qualquer hardware de um conjunto específico (computadores digitais, por exemplo), sem significar, que um software possa ser executado sem estar interligado a um hardware qualquer. Dessa forma, o software é, *em dado sentido*, independente do hardware, a saber, de *um hardware específico*, mas é dependente de hardwares, já que seria obviamente impossível que um programa de computador funcione sem um computador. Essa característica dos estados computacionais se configura como o coração da tese computacionalista, a qual versa que

> Estados mentais são apenas estados computacionais, e estados computacionais são independentes de implementação. Eles têm de ser fisicamente implementados, por garantia, mas não busque a mentalidade na matéria (o *hardware*): é o *software* (o programa de computador) que importa[77]. (HARNAD, 2007, p. 298)

Harnad observa, após delinear mais cuidadosamente a premissa 2, que Searle talvez obtivesse mais sucesso se tivesse deixado claro o que o AQC realmente gostaria de atingir, a saber, a característica de independência em termos de implementação computacionalista. Entretanto, não realizando tal reflexão mais aprofundada e tomando o Computacionalismo de forma bastante trivial, Searle teria acreditado que tal tese se trataria apenas de afirmar que "o cérebro é irrelevante", como se o hardware fosse completamente irrelevante, não compreendendo de que forma a independência de implementação se daria no caso de programas computacionais e, consequentemente, de estados computacionais.

A terceira premissa, segundo Harnad, teria sido menos mal compreendida do que simplesmente incompleta. O autor refez, assim como com as outras premissas, a proposta desta terceira: "Não existe teste empírico, para a presença de estados mentais, mais forte do que a Indistinguibilidade de Turing; por isso o TT é o teste decisivo para uma teoria computacionalista dos estados mentais[78]" (HARNAD, 2007, p. 298). O erro da terceira premissa original (incompleta) é que ela não leva em consideração que poderiam existir testes mais fortes que o TT. Searle acredita que a proposta computacionalista eleva o TT ao patamar máximo de autoridade no que concerne à existência de estados mentais, e que qualquer maquinário que passe no teste recebe o título de possuidor de mentalidade. Não é bem assim, entretanto.

Para ilustrar a capacidade do TT, Harnad nos sugere que imaginemos a simulação de um pato. A primeira simulação (chamá-la-emos de P4) deveria se igualar a um pato real, tanto estrutural quanto funcionalmente: P4 deveria caminhar como um pato, nadar e fazer sons de pato (quack), se parecer com um pato, enfim, parecer ao máximo com um pato de verdade. Ao estudarmos P4, teríamos a mesma riqueza de detalhes referente ao estudo de um pato real, podendo compreender

---

[77] "Mental states are just computational states, and computational states are implementation-independent. They have to be physically implemented, to be sure, but don't look for the mentality in the matter (the hardware): it's the software (the computer program) that matters". (HARNAD, 2007, p. 247)

[78] "There is no stronger empirical test for the presence of mental states than Turing-Indistinguishability; hence the Turing Test is the decisive test for a computationalist theory of mental states". (HARNAD, 2007, p. 298)

como um pato se comporta, já que P4 é isomórfico a um pato real. Esse seria o cenário ideal para um estudo à distância, por exemplo; o nível máximo de precisão que poderíamos atingir. Entretanto, existem níveis inferiores, que requerem menos que isso. Imaginemos uma outra simulação, chamaremos de P3: diferentemente de P4, P3 deteria apenas os aspectos funcionais de um pato orgânico, ou seja, caminharia, nadaria e faria sons de pato (quack) mas não se preocuparia em ter a aparência de um pato, buscando apenas assimilar o comportamento. À medida que P4 seria funcional e estruturalmente isomórfico a um pato, P3 seria apenas funcionalmente isomórfico. Ainda temos um outro nível dessa simulação, chamado de P2, ainda mais limitado: dentre as características funcionais de um pato, P2 buscaria simular apenas os sons emitidos por ele, a saber, o "quack".

Esse nível de simulação pode ser comparado ao nível de simulação que o TT deseja possuir, à medida que possibilita que um computador devidamente programado possa mimetizar a habilidade de comunicação humana. A compreensão de tais níveis de simulação, do macro (funcional + estrutural) ao micro (uma fatia de funcionalidade) dá-nos a noção de que o TT, possuindo apenas o nível de simulação de P2, jamais poderia ser considerado, por alguém razoável, um teste decisivo para atestar a existência de uma consciência, muito menos no sentido da certeza cartesiana dos estados mentais. Somado a isto tem-se também a dificuldade (talvez impossibilidade) de, mesmo através de qualquer dispositivo ou de nossas próprias capacidades, experienciar outras entidades mentais diretamente, "A menos que nós tenhamos uma maneira de realmente *nos tornar* essa outra entidade, e isso parece ser impossível[79]" (HARNAD, 2007, p. 303). Em suma, Searle acreditava que o seu AQC poderia invalidar o TT à medida que demonstrava que este não poderia ser um indicador de estados mentais em, por exemplo, computadores digitais devidamente bem programados; em contrapartida, ironicamente, essa jamais fora a intenção de Turing, sendo o TT assumidamente falível. Ainda que sua ambição fosse bem menos modesta (e o TT fosse considerado um teste comprobatório), segundo Harnad, o AQC poderia contestar apenas uma versão do TT no nível de P2, já que simulações no nível de P3 e P4 jamais poderiam ser refutadas tendo em vista que em nenhum desses casos Searle poderia *ser* o sistema completo, tornando-se tal

[79] "Not unless we have a way of actually *becoming* that other entity, and that appears to be impossible". (HARNAD, 2007, p. 303)

entidade e garantindo (ou não) sua mentalidade com a certeza cartesiana da qual falávamos anteriormente.

3.2.2.2. Rey e a definição de funcionalismo

Ao passo que Harnad focou-se mais no desenvolvimento de uma explanação mais apropriada acerca das intenções de Turing, no que concerne principalmente ao seu artigo de 1950, no qual é apresentado o conceito de Máquina de Turing e do TT, Georges Rey, por sua vez, busca delinear uma definição de funcionalismo que não havia sido abordada por Searle[80], demonstrando que a visão deste em torno dessa teoria da mente seria, no mínimo, incompleta, levando em conta apenas um de seus aspectos, a saber, de acordo com Rey, uma linha de pensamento funcionalista bastante rasa que se baseia em sua maior parte no behaviorismo.

Em seu artigo, *Searle's misunderstanding of Strong AI* (2007), como o leitor pode prever, Rey busca apontar os problemas envolvendo a má compreensão da definição de IA Forte que persiste nos textos searleanos; para tanto, Rey enfoca uma versão específica desta versão forte da IA, nomeada *Teoria Computacional e Representacional do Pensamento* (nos referiremos a ela por TCRP), que por sua vez é considerada por Rey como uma *instância* de uma teoria funcionalista da mente (2007, p. 205) – por este motivo o texto, em sua maior parte, é voltado às desmistificações, esclarecimentos e correções relativas ao pensamento de Searle diante dessa teoria da mente.

Em primeiro lugar, é preciso definir a TCRP e enumerar suas principais características ou reivindicações. Sobre a TCRP, Rey afirma que

> [...] esta é a visão de que atitudes proposicionais (como crença, percepção e preferência) devem ser consideradas como relações computacionais para representações semanticamente valiosas que

[80] Os apontamentos de Rey não se resumem ao artigo *Minds, brains and programs* (1980), no qual Searle apresenta o seu AQC. Neste artigo, Searle não critica explicitamente o funcionalismo, mas ao analisar as objeções ao seu argumento, fundamenta as suas replicas, muitas vezes, em críticas à ideia de que computadores devem ser considerados dotados de consciência pelo fato de *funcionarem* como se possuíssem uma. Searle é mais explícito em sua crítica ao funcionalismo em outras obras, como em *A redescoberta da mente* (2006), onde dedica uma sessão de capítulo para dissertar acerca do que chama de "funcionalismo caixa preta" (2006, p. 62). Tais apontamentos searleanos, na obra supracitada, são muitas vezes alvo de críticas de Rey, inclusive em seu *Searle's misunderstanding's of Strong AI* (2007, p. 201-223).

> são codificadas no cérebro ou em outro hardware daquele que pensa[81]. (2007, p. 203)

Há, entretanto, duas características desta ramificação que se distancia da definição apresentada por Searle em seu texto. Vejamos.

1) A TCRP preocupa-se, principalmente, em construir máquinas que consigam resolver problemas que normalmente requerem inteligência, como por exemplo no caso de jogos, diagnósticos médicos, etc., em vez de se preocupar em produzir dispositivos literalmente inteligentes, que possuam atitudes proposicionais genuínas e uma intencionalidade artificial. A preocupação central dessa proposta de pesquisa são os produtos de inteligência capazes de resolver certos problemas, e não os processos pelos quais tal inteligência é produzida.
2) A TCRP enfoca a intencionalidade, seja ela produzida natural ou artificialmente (através de artefatos), bem como representações de valor semântico. Tal linha de pesquisa tenciona formular uma teoria tanto sobre os processos computacionais que circundam o cérebro quanto sobre seus poderes representacionais, a saber, sua "intencionalidade intrínseca".

Rey deixa claro que a discussão em torno da TCRP não visa legitimar sua verdade ou falsidade, já que esta é uma tese *empírica* (REY, 2007, p. 204) e tais valores só poderiam ser atribuídos com pesquisas mais adiantadas do que as que possuímos hoje. Deve-se buscar (ou não) na TCRP a existência de uma *coerência* de tal tese, ou seja, se a TCRP seria possível como uma explicação dos fenômenos mentais.

A maior distinção da TCRP como teoria funcionalista que pode ser descrita por Rey é aquela diante as teorias behaviorista e do dualismo cartesiano. O funcionalista que engajado na pesquisa da TCRP não cometerá os mesmos erros que carrega o behaviorista e o dualista de propriedade[82], tipicamente cartesiano, por

[81] "[…] this is the view that propositional attitudes (such as believing, noticing, preferring) are to be regarded as computational relations to semantically valuable representations that are encoded in the brain or other hardware o the thinker". (REY, 2007, p. 203)

[82] Estamos considerando "dualistas de propriedade" todo aquele que defende a posição de que, embora haja apenas um tipo de substância existente, a saber, a matéria, tal substância é dividida em duas propriedades, a saber, a mental (responsável pelas crenças, desejos e outros estados intencionais) e a propriedade física (nesse contexto, o cérebro).

identificar os estados mentais preocupando-se com o seu aspecto causal, buscando lidar com essa realidade de maneira não-superficial, preocupando-se não apenas com o comportamento *externo*, mas também com o funcionamento *interno*, até mesmo – e, principalmente – num nível molecular (ou seja, micro, literalmente concernente às moléculas e não de um ponto de vista macro, que pode ser visto a olhos nus), como veremos em seguida. Já mencionamos dois pontos que distanciam a TRCP da definição fornecida por Searle em seu texto; vamos agora enumerar suas características principais:

1) A TRCP, como foi dito, é *molecular*, ou seja, aplicada a subsistemas da mente, e não à mente como um todo. Podemos citar subsistemas como a visão, tomada de decisões e o processamento da linguagem. Os *inputs* e *outputs* em jogo, neste momento, não seriam simplesmente estímulos externos e respostas a tais estímulos, mas em vez disso, informações carregadas num nível micro, entre um e outro subsistema;
2) A TRCP é *ligada* (ou, utilizando o termo cunhado pelo autor, *ancorada*) às várias relações causais presentes no ambiente;
3) A TRCP pode ser considerada um *psicofuncionalismo*, já que suas definições são baseadas em uma psicologia empírica, diferente de teorias que possuem apenas um caráter *a priori* e subjetivo.

Dadas suas principais características, não será difícil concordar o quão marcante é a diferença entre o funcionalismo apresentado por Searle e este que nos foi evidenciado por Rey. Searle em algum momento[83] se referiu a essa teoria como "o funcionalismo da caixa preta", acusando-o de ser bastante limitado no que concerne ao fornecimento de uma definição do que acontece por "dentro" da cabeça, isto é, de que modo os estados mentais se relacionam de maneira causal entre si e com o "mundo exterior" (a saber, a realidade objetiva), praticamente sem fazer distinção entre funcionalismo e behaviorismo. Rey adiciona a forte postura de Searle com relação ao funcionalismo de maneira negativa ao ponto de afirmar que "Se você está inclinado ao funcionalismo, eu acredito que você não precisa de nenhuma refutação, você precisa de ajuda[84]" (SEARLE *apud* REY, 2007, p. 207), desconsiderando

[83] Cf. SEARLE, John R. *The rediscovery of the mind*. MIT Press: Cambridge, 1992, p. 42.
[84] "If you are tempted to functionalism, I believe you do not need refutation, you need help". (SEARLE *apud* REY, 2007, p. 207)

completamente as características mencionadas acima e o aspecto “interno” do funcionalismo.

Trazendo a discussão para nossa temática do quarto chinês, o desfecho final de tal alegoria, como já sabemos, é que apesar do homem dentro do quarto realizar operações que o deixem indistinguível de um autêntico falante chinês, isso não garante que ele compreenda uma palavra de chinês; Searle acredita que o funcionalista discordaria disso, levando em consideração apenas os *inputs* e *outputs*, se satisfazendo apenas com o fato de que, já que as respostas foram entregues corretamente, isso garante o bom funcionamento da sala e isso já seria suficiente. Em contrapartida, supondo que as características apresentadas por Rey sejam verdadeiras, Searle também teria se enganado nesse ponto, posto que o funcionalista examinaria não apenas o comportamento apresentado de maneira externa, a saber, o que acontecia fora da sala (entrada e saída de dados), mas o seu funcionamento *interno*, e após constatar que o que acontecia dentro da sala seria distinto do que ocorre, a um nível molecular, no cérebro de um falante chinês, rapidamente constataria que os processos não são semelhantes, portanto não defenderia aquilo que Searle tanto critica.

Um outro ponto que deve ser trazido à tona no texto de Rey é a maneira pela qual a linguagem é abordada pela TRCP (e, consequentemente, pelo funcionalismo). Mais uma vez, em contraste com o que Searle poderia imaginar, a TRCP não acredita que seja possível aprender um novo idioma apenas decorando um livro com alguns símbolos ou um manual básico de expressões idiomáticas. Segundo Rey,

> [...] é melhor que haja vários tipos de conexões causais entre símbolos internos e fenômenos externos. Indiscutivelmente, compreender uma linguagem envolve estar apto a relacionar símbolos de uma língua a pelo menos *algumas* percepções, crenças, desejos, e alguns tipos de disposições de comportamento: por exemplo, *ceteris paribus*, aplicar a palavra chinesa para neve a amostras reais de neve, responder com a palavra chinesa referente a branco quando você for interrogado sobre a cor da neve[85]” (2007, p. 208).

[85] “[...] there had be better be various kinds of causal connections between internal symbols and external phenomena. Arguably, understanding a language involves being able to relate the symbols of the language to at least *some* perceptions, beliefs, desires, and some sort of dispositions to behave: for example, *ceteris paribus*, to apply the Chinese word for snow to actual samples of snow, to respond with the Chinese word for white when you take yourself to have been asked the color of some snow”. (REY, 2007, p. 208)

Ao apresentar essa concepção de linguagem, Rey decide delinear brevemente uma versão alterada do quarto chinês, à luz de uma concepção funcionalista e da TRCP. Em linhas gerais, um programa a fim de mimetizar a mente deveria possuir, pelo menos subsistemas como a percepção, raciocínio, memória, tomada de decisão, processamento de linguagem e controle motor, além de uma bateria de "subprogramas" para lidar com cada subsistema referido. Levando esses requisitos para a ambientação do quarto chinês, seria necessário todo um conjunto de programas para executar tais operações e, como seria impossível para que uma pessoa só realizasse todas essas tarefas paralelamente, várias pessoas seriam necessárias para tal empreitada, cada uma executando um programa diferente. Tal projeto é intitulado 'Mente Modesta', sendo uma simplificação radical do tipo de riqueza computacional que seria necessária para simular um falante chinês. Dessa maneira, não apenas o comportamento externo seria mimetizado, mas também o processamento interno. Ainda assim, afirma, Rey, Searle não duvida que seja possível construir um dispositivo capaz de tal feito (como de fato é possível); o seu problema, por outro lado, pode ser resumido na afirmação de que, mesmo que seja possível algo desse tipo, ainda assim a inteligência genuína não poderia ser assegurada. Entretanto, "[...] uma vez que nós mimetizamos a verdadeira forma através da qual as pessoas produzem seu comportamento inteligente, é difícil enxergar como essa objeção ainda pode ser levantada[86]" (REY, 2007, p. 212).

Podemos concluir as considerações de Rey apontando, em caráter de síntese, que a TRCP possui caráter de um funcionalismo molecular e ancorado, buscando definir estados mentais em termos de vários subsistemas e de suas relações causais entre si e externamente; essa teoria é um *programa de pesquisa*, dependendo de outras áreas como a psicologia, biologia, medicina e, por causa disso, acredita-se (os defensores de tal teoria, obviamente) que tal programa pode gerar contribuições significantes para uma teoria da compreensão humana.

A compreensão humana, foco principal das preocupações de Searle ao desenvolver o seu AQC, ainda não foi totalmente explicada pela ciência ou pela filosofia: a mente ainda é um mistério e tanto, ainda que áreas como a neurobiologia continuem avançando paulatinamente ao longo dos anos. Não possuímos uma

[86] "[…] *once we have mimicked the actual way people do produce their intelligent behavior,* it's hard to see how this objection can still be raised". (REY, 2007, p. 212)

resposta final para o problema da consciência, da mente, da compreensão e do problema mente-corpo. Entretanto, muitos cientistas e filósofos buscam dar suas contribuições para áreas que tematizam esses problemas. Searle, como sabemos é um deles. A partir de então nos cabe demonstrar de que forma Searle pôde dar sua contribuição à área da filosofia da mente, ou melhor colocando, à área da filosofia da mente voltada às investigações concernentes à Inteligência Artificial e/ou ao modelo computacional de mente.

4. O NATURALISMO BIOLÓGICO: UM DESFECHO

Durante os capítulos que se seguiram, nosso esforço foi aplicado na apresentação da discussão em torno do AQC: do que se trata o AQC, quais são seus pressupostos básicos, o que está explícito e o que não está na argumentação de Searle, o que exatamente o autor visava criticar; mais tarde, apontamos e discutimos as críticas a esse argumento, buscando apresentar de vários ângulos os pontos fracos do texto searleano. Nossos esforços voltaram-se a esse fim tendo em vista localizar a crítica de Searle, através do AQC, na área da filosofia da mente, para que, finalmente, pudéssemos extrair o que havíamos antes proposto como escopo de nosso trabalho: a contribuição de Searle para a filosofia da mente[87], no que diz respeito às discussões em torno da IA. Dedicaremos esse capítulo ao trato deste nosso objetivo, buscando apresentar a abordagem searleana em geral (leia-se o conjunto de ideias voltadas à problemática da Inteligência Artificial, não somente o AQC) que acreditamos configurar a sua real contribuição para a Filosofia da IA e, consequentemente, para a Filosofia da Mente como um todo.

A pergunta: "*qual a contribuição de Searle para a área da Filosofia da Mente, mais especificamente no tocante à Inteligência Artificial?"* não pode ser respondida de maneira simples, e talvez não haja apenas uma resposta, mas várias. Podemos dizer que o Argumento do Quarto Chinês está proximamente ligado (e apresentaremos como) a um importante posicionamento de Searle, conhecido como a teoria do *naturalismo biológico*. À primeira vista, tal teoria representa mais uma no quadro de ideias e concepções que buscam explicar, simplificar ou até mesmo eliminar, o problema mente-corpo. Entretanto, se tomarmos a teoria tendo em vista o pano de fundo da IA, veremos que ela desempenha um papel importante no que concerne ao AQC, sustentando todas as implicações e críticas que o argumento

[87] Muitos filósofos reconhecem a contribuição de Searle para essa área. Podemos citar como exemplo Preston, que afirma, sobre o AQC searleano, que ele "[…] prejudica a autoimagem oficial da Inteligência Artificial (IA), um dos supostos fundamentos de boa parte da ciência cognitiva contemporânea. Ele também pode muito bem ser o argumento mais conhecido na filosofia contemporânea" (2007, p. 2). Bringsjord & Noel acrescentam: "O Argumento do Quarto Chinês de John Searle é indiscutivelmente o maior polarizador filosófico do século XX. [...] de acordo com o passar dos anos, parece mais e mais como se seu argumento estivesse se dirigindo à imortalidade filosófica" (2007, p. 144). Mesmo um de seus críticos, Stevan Harnad, reconhece sua importância: "[...] o artigo de Searle tornou-se o artigo direcionado mais influente da BBS (e ainda é, até a presente data), da mesma forma que é um clássico na ciência cognitiva" (2007, p. 205). É possível também identificar a importância dos posicionamentos de Searle, a partir do AQC, em livros-texto sobre IA, a exemplo da obra de Guilherme Bittencourt (2006, p. 31). Além disso, o artigo do matemático Alan Turing (1950) dificilmente é desvinculado, nos dias atuais, da crítica searleana ao mesmo.

evidencia contra as propostas da IA Forte, do computacionalismo e de tudo aquilo que tratamos nos capítulos anteriores. Diante disso, se houvesse uma resposta simples a ser dada à pergunta do início deste parágrafo, esta seria: a teoria do naturalismo biológico.

A partir de então o leitor pode vir a se perguntar: “Se a teoria do naturalismo biológico representa a contribuição de Searle, então como a mesma pode vir a alterar, melhorar ou revisar o que a Inteligência Artificial propôs até então? Em outras palavras, como, efetivamente, esta teoria contribui?”. A fim de responder a estes questionamentos, vamos, então, seccionar este capítulo em duas seções: em primeiro lugar, enumeraremos as contribuições da teoria do naturalismo biológico para o quadro geral da IA e em segundo lugar, analisaremos o conceito de IA tendo em vista tudo o que foi discutido até então.

4.1. O naturalismo biológico: quatro características

Antes de tudo, vamos relembrar do que se trata a teoria do naturalismo biológico: o naturalismo biológico é aquela teoria que afirma que a consciência não é matéria e nem pode ser reduzida à matéria, ao mesmo tempo que não se trata de uma substância distinta e apartada do material. Ao contrário disso, a consciência *acontece* no organismo do ser consciente (a saber, no cérebro) e é *causada* no mesmo organismo, sendo, portanto, resultado de um processo, sem poder ser reduzida aos elementos que tornaram possível esse processo. (SEARLE, 2006, p. 26). Em outras palavras, a consciência está para o cérebro da mesma forma que a liquidez está para a água ou que a digestão está para o aparelho digestivo. De acordo com Searle,

> [...] o fato de uma característica ser física não implica que não seja mental. Revisando Descartes, por enquanto poderíamos dizer não somente ‘penso, logo existo’, e ‘sou um ser pensante’, mas também *sou um ser pensante, portanto sou um ser* físico. (2006, p. 26)

Essas afirmações compreendem o esforço de Searle para deixar claro para o leitor que sua tese não se trata de mais uma que pode ser categorizada como dualista de propriedades: a consciência é, sim, uma propriedade do cérebro, mas não é de mesmo tipo que o dualista de propriedades imagina: para este dualista, esta

propriedade especial do cérebro seria *epifenomenal*, ou seja, estaria *acima*, de forma isolada, do que acontece no cérebro. O que isso quer dizer? Tais propriedades seriam *causadas* pelo cérebro, mas por sua vez não possuiriam quaisquer poderes causais sobre ele (CHURCHLAND, 2004, p. 31). Vejamos agora, de forma mais ampla, algumas características da teoria do naturalismo biológico.

4.1.1. Dualismo, materialismo e o meio alternativo

O próprio Searle afirma (2006, p. 24) que sua teoria já o colocou em lugares indesejados: por alguns foi catalogado como dualista de propriedades, já por outros foi até descrito como defensor de uma filosofia mística. Por quê? A causa disto seria, talvez, por Searle não situar exatamente onde se "encontra" a consciência: se estaria num plano material ou num plano efêmero. Dessa forma, acreditou-se por vezes que o autor defendia que a consciência seria uma propriedade emergente e especial da matéria, sendo caracterizado como dualista de propriedades.

Em *A redescoberta da mente* (2006), Searle expõe de forma muito detalhada as causas dessa confusão: por muito tempo cultuou-se uma falsa dicotomia, em que só existia duas classes de filósofo da mente: materialista ou dualista. Mesmo no senso comum, a divisão é clara: o mental é sempre oposto ao físico, assim como o corpo é à mente, a matéria é ao espírito, dentre outros. É como se não houvesse um fenômeno que pudesse transitar no meio disso tudo: ou uma coisa, ou outra. Indo um pouco mais além, há uma facilidade em se adjetivar um pensador de "cartesiano" puramente por ele aceitar a existência do mental de forma similar a Descartes. Diferentemente de alguns pensadores (conhecidos como materialistas eliminacionistas), o mental não é algo que ocupe uma esfera privilegiada e apartada de tudo o que chamamos de "objetivo" no mundo. O mental desempenha um papel causal em nossas vidas cotidianas. No mundo há objetos e coisas que independem desse mental, de estados mentais; a exemplo disso temos a matéria: ela é matéria independentemente de qualquer coisa. Entretanto, utilizando os exemplos de Searle (2006, p. 40), "problemas na balança de pagamento, sentenças não-gramaticais, razões para suspeitar da lógica modal, minha habilidade para esquiar", são elucidados por nós de forma intencional, num emaranhado de crenças e desejos; muitas vezes são conceitos, que partem da consciência e que afetam o cotidiano "físico". Não é preciso ser cartesiano para compreender isso. Ainda assim, segundo

Searle “Em parte, sem dúvida, simplesmente por negligência intelectual (ou talvez ainda pior) da parte dos comentadores”, faz-se questão de que sejam exibidas apenas duas categorias: o dualismo e o materialismo, causando a falsa dicotomia que mencionamos anteriormente.

Não que isso não tenha ficado claro ainda, mas é importante voltar a frisar que a consciência é uma propriedade *emergente* ou de *nível superior* do cérebro. Isso significa que ela não é caracterizada por uma relação estrita entre neurônio A e neurônio B, num nível micro, mas ao contrário, é “encontrada”, por assim dizer, no nível macro do cérebro, chamado por Searle de nível superior. Em outras palavras, a consciência não pode ser resumida, ou melhor, *reduzida* a essa relação entre os neurônios. A fim de resumir e arrematar o que afirmamos até agora neste tópico, vale a pena citar Searle:

> A consciência é uma propriedade mental, e portanto física, do cérebro, no sentido em que a liquidez é uma propriedade de sistemas de moléculas. [...] O fato de uma característica ser mental não implica que não seja física; o fato de uma característica ser física não implica que não seja mental. Revisando Descartes, por enquanto poderíamos dizer não somente ‘penso, logo existo’ e ‘sou um ser pensante’, mas também *sou um ser pensante, portanto sou um ser físico*. (2006, p. 26)

4.1.2. A realidade é objetiva: isso pode ser questionado

Além do cartesianismo e da falsa dicotomia, uma outra característica que posiciona a teoria searleana em um mal local (mística, dualismo de propriedades etc.) é o costume decorrente desde o século XVII de se acreditar e aceitar facilmente a pressuposição de que a realidade é objetiva. Segundo Searle, tal assertiva, apesar de amplamente aceita, é *falsa*: “Essa assunção mostrou-se útil para nós de muitas maneiras, mas é obviamente falsa, como revela um momento de reflexão sobre os estados subjetivos próprios de qualquer pessoa” (2006, p. 28). Talvez por medo de se cair nas graças do cartesianismo, pesquisadores de diversas áreas acreditaram que seria mais ‘científico’ assumir que a mente, como parte de uma realidade objetiva, deveria ser estudada objetivamente a partir de eventos que pudessem ser observados por qualquer um. Com isso, o foco não seria mais a experiência pessoal e subjetiva da primeira pessoa, mas sim o comportamento observável. Sagazmente Searle pontua (2006, p. 28) que as perguntas foram feitas de maneira equivocada:

em vez de perguntarmos “o que é ter um desejo?”, perguntávamos “como podemos *atribuir* a uma pessoa ou sistema, estados como crenças, desejos, tudo isso através de uma metodologia que observe o comportamento e impulsos externos?”.

Existe, em meio a tudo isso, um equívoco entre *epistemologia*, *ontologia* e *causação*. A pergunta que compete à epistemologia claramente é aquela relacionada ao *modus operandi* científico, ou seja, o de observador em terceira pessoa. Quando essa pergunta é feita em detrimento da pergunta ontológica (ligada ao ponto de vista da primeira pessoa), uma porta é aberta para que concebamos que não apenas seres humanos e outros animais possuem mentes/consciência, mas também que qualquer coisa que se comporta *como se* possuísse, por exemplo, um computador, de fato possuísse. É importante lembrar que todos os estados conscientes são estados *de alguém*, e portanto o ponto de vista de primeira pessoa jamais deverá ser descartado.

Searle lança mão do conceito de ontologia de uma maneira diferente da tradicional, de forma mais “frouxa”, por assim dizer. Em *A redescoberta da mente* (2006, p. 31), o autor limita-se apenas a expor o que ele chama de “a pergunta” que compete a cada abordagem (epistemologia, ontologia e causação). A intenção do autor ao utilizar o conceito de ontologia é apenas para apontar a “visão da primeira pessoa”, tomando como ontológico aquilo que concerne ao ser (não o Ser aristotélico, mas o ser no sentido corriqueiro e de senso comum), como na pergunta: “o que *é* isto?”. A causação, por sua vez, é identificada por Searle principalmente pelas relações causais entre os estados mentais e com relação ao mundo. A distinção entre epistemologia, ontologia e causação evita que alguns equívocos sejam cometidos, como veremos no próximo tópico.

4.1.3. Epistemologia, ontologia e causação

Em linhas gerais, a epistemologia, ontologia e a causação possuem diferentes formas de abordar um mesmo conteúdo. De acordo com Searle (2006, p. 31) podemos concluir inicialmente que a epistemologia possui uma abordagem ligada à preocupação de como conhecemos algo: “como obtemos conhecimento disto?” é uma pergunta epistemológica. A ontologia, por sua vez busca conhecer o que é: “o que é isto?” é uma pergunta da ontologia. A causação, por sua vez, busca

conhecer o que causa certos fenômenos e coisas no mundo: “o que isto causa?” é a pergunta da causação.

Como mencionamos no tópico anterior, é recorrente que, devido a uma dada herança filosófica, suponhamos que haja apenas duas teorias, a saber, a dualista e a materialista, ou duas categorias, o material e o mental, e por medo de recorrermos ao cartesianismo, ‘cientifiquemos’ completamente a abordagem da consciência, resumindo-a à observação em terceira pessoa. Nesse sentido, a única pergunta que está sendo feita é a epistemológica: “como obtemos conhecimento disto?”, e respondemos a essa pergunta através da busca pelo comportamento e pela relação de *input* e *output*. A problemática em tratar a consciência apenas através desta via é que acabamos por trata-la

> [...] independentemente da consciência, isto é, trata-la unicamente a partir de um ponto de vista de terceira pessoa, e isto leva à concepção de que a consciência como tal, como eventos fenomênicos “internos”, “privados”, não existe realmente. (SEARLE, 2006, p. 33)

Mesmo que a expressão pareça confusa, Searle afirma que é possível tratar a consciência *independentemente* da consciência, ao tratarmo-la deixando de levar em consideração todos os seus aspectos subjetivos, um deles sendo os *qualia*. Como a única metodologia em voga é aquela que envolve o ponto de vista de uma terceira pessoa, não é possível detectar e diferenciar os aspectos qualitativos dos estados mentais. Esse é apontado como um dos maiores problemas do behaviorismo filosófico. Ao subordinar a ontologia à epistemologia e detectar alguns problemas (como o dos *qualia*), tenta-se então contornar a situação voltando-se para aspectos causais, buscando-se compreender não apenas o que ocorre no “exterior”, mas agora como se daria o funcionamento “interno” do mental, numa tentativa de substituir isto por tudo o que é subjetivo: essa é a empreitada do funcionalismo. A partir daí, como se assume facilmente que os estados mentais podem ser definidos apenas por suas relações causais, não é difícil se chegar à conclusão de que estados mentais são *apenas* estados computacionais (proposta do computacionalismo, acolhida pela IA Forte (SEARLE, 2006, p. 35)).

À primeira vista, parece que essa é a única resposta ao problema das outras mentes, por exemplo. “Como saber se existem outros seres pensantes por aí?”, alguém poderia se perguntar, ao passo que outro iria responder “Ora, através da

observação de seu comportamento! Aliás, não apenas isto: certifique-se que o objeto se comporte como você e que funcione semelhantemente a você. Se isso acontecer, provavelmente será igualmente consciente". E mais uma vez chegamos à falsa dicotomia: introspecção de um lado, comportamento de outro. Searle, perspicazmente, sugere mais uma vez o meio alternativo:

> Naquilo que diz respeito ao conhecimento de outras mentes, o comportamento *sozinho* não tem interesse para nós; é antes *a combinação do comportamento com o conhecimento dos sustentáculos causais do comportamento* que forma a base de nosso conhecimento. (SEARLE, 2006, p. 36).

Essa sugestão dá início à resposta do problema das outras mentes, o qual era antigamente respondido simplesmente pelo comportamento. De acordo com o autor (SEARLE, 2006, p. 109), não é suficiente apenas que algo ou alguém se comporte de maneira tal que se assemelha ao comportamento consciente, mas é necessário o conjunto que envolve esse fato somado a mais um outro: que o fundamento causal do comportamento fisiológico do indivíduo seja semelhante ao de quem é consciente.

Esses dois pontos, tomados de maneira isolada, não são suficientes para convencer-nos de que um indivíduo possui estados mentais conscientes; sua *conexão*, por outro lado, que pode ser lida como "o comportamento apropriado, que tem a *causação* apropriada na fisiologia subjacente" (SEARLE, 2006, p. 110), é relevante para a descoberta dos estados mentais e/ou das outras mentes. Vale salientar: Searle não afirma que o comportamento não importa em nenhuma instância; obviamente, no cotidiano, o comportamento é o que nos faz agir como agimos naturalmente e é por supor, através de senso comum, que determinado comportamento implica em determinada natureza de coisas, que seres humanos e outros animais conseguiram sobreviver a intempéries; por exemplo, podemos reconhecer um cachorro ou outro animal (inclusive predadores) a partir de algumas características gerais acerca de sua aparência e comportamento; nós supomos, rapidamente, que aquele organismo é tal animal, pela forma como se comporta e pela similaridade com relação a outros animais, de mesma espécie, que conhecemos. O que Searle procura deixar claro é que, no que diz respeito à *ontologia da mente*, o comportamento *isoladamente* é irrelevante.

Ainda sobre a importância do comportamento, Searle (2006, p. 103) pontua até onde ele (o comportamento) pode ser considerado como fator altamente relevante. Mais uma vez, a distinção clara entre epistemologia, causação e ontologia é requisitada. *Ontologicamente falando*, o comportamento seria irrelevante para a existência de fenômenos mentais conscientes; imaginemos um ser humano enfermo, sem possuir domínio sob seus movimentos, sem conseguir mexer um dedo sequer. Ele não responderia a estímulos e não pareceria sequer perceber a presença de estímulos. Entretanto, "interiormente", suas faculdades mentais poderiam estar intactas. Há, de fato, uma ligação entre comportamento e consciência, onde esta pode *causar* aquele, mas esta não se trata de uma relação de interdependência; a essas observações Searle (2006, p. 104) dá o nome de "princípio da independência de consciência e comportamento".

O princípio supracitado, assim como outros pontos que iremos evidenciar em breve, é uma das mais importantes contribuições que o naturalismo biológico pode acrescentar à filosofia da mente e à IA. Ele abre portas para que mais meios alternativos sejam iniciados, e que ao instanciar um programa de computador, por exemplo, tenha-se em vista não somente os aspectos comportamentais/funcionais e causais de maneira isolada, mas também os ontológicos. Se a IA (no caso, a tendência mais forte) será capaz de se enveredar por um caminho assim sem trair seus objetivos iniciais, é algo que iremos discutir mais a fundo na segunda seção deste capítulo.

4.1.4. Intencionalidade: intrínseca, derivada e "como-se"

O conceito de intencionalidade está presente, explícita ou implicitamente na maior parte do discurso searleano, além de ser um conceito fundamental para a discussão que envolve a Inteligência Artificial. Searle define provisoriamente o termo intencionalidade como *direcionalidade*, no sentido em que esta seria a faculdade pela qual a mente humana relaciona-se com vários estados de coisas no mundo. Nas palavras do próprio filósofo, "[...] Intencionalidade é aquela propriedade de muitos estados e eventos mentais pela qual estes são dirigidos para, ou acerca de, objetos e estados de coisas no mundo" (SEARLE, 2007, p.1). Até este ponto da conceituação do termo, o filósofo segue a existente tradição filosófica referente à intencionalidade. Entretanto, há uma ambivalência conceitual, graças à teoria

searleana, que devemos tomar nota: o filósofo afirma (SEARLE, 2007, p. 6-18) ainda que a intencionalidade não apenas é algo que se direciona ao mundo, mas que ela é a *representação* (que nesse contexto deverá ser lida como requerimento) do ato de fala - Searle apresenta a intencionalidade como anterior ao ato de fala, ou seja, o ato de fala depende da Intencionalidade, relacionando-se causalmente com ela.

Como pudemos verificar, conceitos de *inteligência, consciência* e *intencionalidade* (apesar destes dois últimos estarem intrinsecamente ligados, não podemos determinar uma sinonímia neste caso, por existir estados intencionais que não são conscientes) são distintos e desempenham papéis diferentes na problemática que viemos abordando desde então. Por este motivo, devemos perscrutar as três formas através das quais se apresenta a intencionalidade, sendo uma delas falsa, como veremos a seguir.

O primeiro caso que podemos mencionar é a *intencionalidade intrínseca*, que pode ser representada através de uma sentença verdadeira deste tipo:

1) Estou severamente inclinado a acreditar que está chovendo pois verifico, através da vista da minha janela, que está caindo água dos céus.

É possível afirmar que existe uma intencionalidade intrínseca a partir de 1 pelo fato de que, se a sentença for realmente verdadeira, deve existir um estado intencional, a saber, uma crença, que atribuiremos ao sujeito da frase. O sujeito, através de um estado mental (crença), direciona-se ao mundo e busca uma direção de ajuste a partir dele, buscando adquirir o conhecimento da existência de chuva naquele dado momento, para que então sua crença seja confirmada como verdadeira. Esse é um exemplo típico de estado intencional e de intencionalidade intrínseca. Podemos ainda acrescentar a isso a seguinte informação:

> Intencionalidade intrínseca é um fenômeno que seres humanos e determinados outros animais têm como parte de sua natureza biológica. Não é uma questão de como são tratados, ou como se concebem a si mesmos, ou de que forma preferem descrever-se a si mesmos. É simplesmente um fato evidente em tais animais que, por exemplo, algumas vezes fiquem com *sede* ou *fome*, *vejam* coisas, *temam* coisas etc. (SEARLE, 2006, p. 118).

O segundo tipo de intencionalidade, que ainda assim é legítimo, é chamado de *intencionalidade derivada*, podendo ser ilustrada a partir da seguinte sentença:

2) Em japonês, “*Watashi wa anata o shinjite iru*” significa “eu acredito em você”.

Dizemos que a sentença acima possui uma intencionalidade derivada pelo fato de que para os japoneses, a sentença em japonês realmente significa algo, a saber, um estado intencional que se direciona ao mundo. É um estado intencional legítimo, mas para os falantes do português, tal estado intencional é transmitido na sentença de forma derivada. Em outras palavras, a expressão linguística carrega, de fato, intencionalidade, mas em 2, ela carrega a intencionalidade dos falantes japoneses.

O terceiro tipo, por sua vez, considerado o mais problemático, principalmente por encontrarmos evidências dele na discussão em torno da IA Forte, é o que Searle chama de “como-se”. Esse não é um tipo genuíno de intencionalidade, e como o nome sugere, é apenas *como se* houvesse intencionalidade, quando na verdade não há (SEARLE, 2006, p. 116). Vamos representá-lo pela seguinte sentença:

3) O meu anti-vírus acredita que meu computador está com um vírus.

É muito comum atribuir intencionalidade a sistemas que na verdade não a possuem, como termostatos, computadores, calculadoras e outros objetos cotidianos. Pensamos em tais objetos como extensões de nossa própria intencionalidade, então é comum ouvirmos que “este computador pensa mais rápido do que aquele, por ter mais memória” ou que “o meu GPS sabe onde eu estou”. Sobre esse caso, Searle (2006, p. 121) ainda afirma que:

> Qualquer tentativa de rejeitar a distinção entre intencionalidade intrínseca e *como-se* enfrenta uma *reductio ad absurdum* geral. Se você rejeita a distinção, resulta que tudo no universo tem intencionalidade. Tudo no universo segue leis da natureza e, por esta razão, tudo se comporta dentro de um determinado grau de regularidade e, por esta razão tudo se comporta *como-se* estivesse seguindo uma regra, tentando executar um determinado projeto, atuando de conformidade com determinados desejos etc.

A distinção dos diversos tipos de intencionalidade, além da aparente intencionalidade que tratamos por último, é um dos pilares da crítica searleana à IA Forte, como já havíamos mencionado anteriormente em nosso texto. A ligação do conceito de intencionalidade à estrutura fisiológica de certos organismos, a saber,

seres humanos e alguns outros animais, é mais uma contribuição searleana que deve ser pontuada: em Searle, o conceito de intencionalidade é reformado, e deixa de significar tão somente *direcionalidade*, ficando raízes em sua teoria do naturalismo biológico.

4.1.5. Um resumo

Finalizamos, dessa maneira, as quatro características do naturalismo biológico. Podemos acrescentar ainda algumas notas com relação ao misticismo e ao dualismo de propriedades, ao qual Searle é, segundo o próprio autor (2006, p. 24), relacionado em detrimento das suas afirmações sobre as características emergentes da consciência. Em que sentido podemos afirmar que a consciência é emergente? Ela é emergente enquanto resultado de um processo ocorre no e é causado pelo cérebro, através das atividades dos neurônios. Como relacionamos anteriormente, a consciência está para o cérebro assim como a liquidez está para a água. A consciência não pode ser reduzida à mera atividade neuronal. A este tipo de característica emergente, Searle dá o nome de *concepção de emergência causal*, ou "emergente 1" (2006, p. 162). Há, por outro lado, um outro tipo de natureza emergente muito mais radical, onde não apenas se detecta uma emergência de tipo 1, mas também a impossibilidade de explicação de tal fenômeno a partir das interações causais dos elementos que compõem aquilo do qual o fenômeno emerge. Tomemos como exemplo a própria consciência: ela emerge da interação entre os neurônios; os neurônios, nesse caso, são os elementos (a grosso modo) do cérebro, a partir de onde emerge a consciência. O fenômeno emergente do tipo 2, mais radical, não pode ser explicado tendo em vista nenhum desses elementos, e é como se ganhasse vida própria, causando coisas que não podem ser explicadas. Esse tipo de fenômeno, emergente 2, pode ser relacionado a uma mística e a um dualismo de propriedades, mas não o do tipo 1, o qual Searle está inclinado a defender.

Diante dos quatro pontos que expusemos, resta o questionamento: de que modo essas características compreendem a contribuição da tese do naturalismo biológico? Em primeiro lugar, o apontamento de uma terceira via, entre materialismo e dualismo sugere também que a filosofia da mente e, se formos adotar o ponto de vista searleano, também a IA em seu sentido forte, repense os suas categorias: até que ponto a rejeição do subjetivo é viável? Searle sugere, até mesmo, que dualistas

e monistas podem estar olhando para o lado errado, quando iniciaram a contagem de quantas propriedades ou substâncias podem existir, e acabaram criando um outro tipo de dualismo, que o autor chama de “dualismo de conceitos”, o qual consiste no ponto de vista que instaura a falsa dicotomia de que algo físico implica no não-mental e por sua vez, mental implica em não-físico. Tanto materialistas quanto dualistas concordam com essa oposição e, diante disso, até mesmo os materialistas poderiam ser associados aos dualistas, o que compele Searle a afirmar que “O materialismo é, portanto, em certo sentido, a mais fina flor do dualismo” (2006, p. 42).

O nosso segundo ponto acerca do naturalismo biológico carrega sua contribuição removendo o foco da suposta objetividade da realidade e da confiança em experimentos voltados apenas para a observação em terceira pessoa, relembrando a importância da subjetividade e das experiências pessoais e subjetivas de cada indivíduo consciente, já que todo e qualquer estado mental, é estado mental *de alguém*. A contribuição que o nosso segundo ponto carrega é ainda mais palpável do que apresentamos anteriormente, já que pode ser considerada como uma resposta a um dos pontos argumentativos de Turing em seu famoso texto *Computing machinery and intelligence* (1950). Turing, assim como Searle, após apresentar o coração de sua tese, definindo o que seria o que hoje chamamos de Teste de Turing (chamado pelo matemático de *jogo da imitação*), trata de apresentar uma série de objeções à sua argumentação, buscando refutá-las uma a uma, de maneira bem similar a Searle. Uma dessas objeções deve chamar nossa atenção de maneira diferenciada pela sua similaridade com o pensamento searleano; é chamada *o argumento da consciência*:

> Este argumento está muito bem expresso no discurso “Lister” do Prof. Jefferson, de 1949, de onde transcrevo: “Somente quando uma máquina puder escrever sonetos ou compor concertos como resultado de pensamentos ou emoções sentidas, e não por via de ocorrência causal de símbolos, é que concordaríamos em que a máquina é igual ao cérebro – isto é, que não apenas os escreveu ou compôs como também sabia que os escrevera. Nenhum mecanismo poderia experimentar (e não meramente assinalar de modo artificial, por meio de uma engenhoca fácil) prazer pelos seus êxitos, tristeza quando suas válvulas queimam, deleite ante a lisonja; sentir-se infeliz por causa de seus erros, encantar-se com o sexo, ficar irritado

> ou deprimido por não poder alcançar o que deseja"[88]. (TURING, 1950, p. 443)

Turing, a fim de objetar a favor de sua tese, constrói o seu discurso com base numa defesa voltada apenas para um ponto de vista epistemológico, dependente única e simplesmente de observação comportamental: o mesmo afirma (1950, p. 443) que a única forma de saber que a máquina pensa é *sendo* a máquina, e como isso não é possível, a única via através da qual se tem acesso a tal conhecimento é através da observação do comportamento. Vimos, em *4.1.2.*, que o comportamento *por si só* não é suficiente para determinar a existência de uma consciência.

O terceiro ponto, por sua vez, complementa aquele segundo, diferenciando uma abordagem ontológica (primeira pessoa) de uma epistemológica (terceira pessoa), sugerindo uma fórmula mais completa para que o problema mente-corpo seja solucionado: observando o comportamento em *conexão* com as interações causais que ocorrem na fisiologia do organismo em questão, responsáveis por fenômenos como a consciência e a intencionalidade.

Por fim, o quarto ponto, talvez o mais explícito na argumentação presente no Argumento do Quarto Chinês, evidencia a confusão que se instaura quando confundimos intencionalidade intrínseca com algo que funciona *como-se* fosse intencionalidade, além de enfatizar a importância desta para toda e qualquer proposta de duplicação de consciência: não basta apenas replicar amostras de inteligência, é preciso explorar as consequências de uma mente consciente de forma abrangente, sem que a intencionalidade seja marginalizada.

4.2. O conceito de Inteligência Artificial: uma metamorfose

De volta a 1950, ao final de seu artigo, Turing afirma algo que preocuparia Searle, cerca de 30 anos depois de escrito:

---

[88] "This argument is very, well expressed in Professor Jefferson's Lister Oration for 1949, from which I quote. Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants'". (TURING, 1950, p. 443)

> É de esperar que as máquinas acabem por competir com o homem em todos os campos puramente intelectuais. Quais, porém, os melhores para começar? Mesmo esta é uma decisão difícil. Muitas pessoas acham que uma atividade bastante abstrata, como o jogo de xadrez, seria o melhor. Pode-se também sustentar que o mais conveniente é prover as máquinas dos melhores órgãos sensoriais que o dinheiro possa comprar, e **ensiná-las a compreender e falar inglês**. Tal processo poderia acompanhar o do ensino normal de uma criança. Coisas seriam apontadas e nomeadas, etc. Mais uma vez, não sei qual a resposta certa, mas penso que ambos os enfoques deveriam ser tentados[89]. (TURING, 1950, p. 460, grifo nosso)

Através do Argumento do Quarto Chinês e de sua tese do naturalismo biológico, Searle identifica precisamente os problemas nos planos de Turing, principalmente no que concerne à compreensão e ao aprendizado. Vimos, no ponto 4, a diferença entre intencionalidade intrínseca e a "intencionalidade" *como-se*. Ficou claro que no caso da citação acima, é atribuída intencionalidade ao computador, mas algo assim não seria genuíno, segundo Searle. Além disso, o AQC aponta os problemas na "compreensão" de uma língua por parte de um computador, apenas em virtude da manipulação de símbolos: há apenas sintaxe, deixando-se a semântica de lado. Não há, de fato, compreensão ali, apenas cálculos.

Em 1950, aquele que hoje é considerado por muitos o pai da computação[90], idealizou as características de um computador digital da forma descrita anteriormente. Mas desde então, será possível afirmar que as pretensões da Inteligência Artificial mudaram? Searle elabora o seu AQC em 1980, mas 15 anos depois Fodor fornece, em entrevista, várias informações pertinentes a um conceito

---

[89] "We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried". (TURING, 1950, p. 460)

[90] Jack Copeland, em seu texto intitulado *The Essential Turing- Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma* (2004), cita um depoimento de Stanley Frankel acerca do papel desempenhado por Turing no contexto da computação, como segue: "[...]Muitas pessoas têm aclamado von Neumann como o "pai do computador" (no sentido moderno do termo) mas tenho certeza de que ele nunca teria cometido esse erro. Ele poderia muito bem ser chamado de a parteira, talvez, mas ele enfatizou firmemente para mim, e para os outros, eu tenho certeza, que a concepção fundamental pertence ao Turing – na medida em que não foi antecipado por Babbage, Lovelace, dentre outros. No meu ponto de vista, o papel essencial de von Neumann estava em tornar o mundo ciente desses conceitos fundamentais introduzidas por Turing e do trabalho de desenvolvimento realizado na escola de Moore e em outros lugares" (COPELAND, 2004, p. 22).

que atravessou décadas intrigando tantos filósofos. *Speaking minds* (1995), a coletânea de entrevistas elaborada por Peter Baumgartner e Sabine Payr reaviva a discussão em torno da intencionalidade, computacionalismo, inteligência e computadores. Jerry Fodor é um dos entrevistados, e a partir do seu depoimento podemos constatar que houve, sim, um processo metamórfico durante as décadas, que tornou a Inteligência Artificial algo diferente, ou, melhor colocando, que fez com que compreendêssemos a IA de maneira diferente.

Há vários aspectos do depoimento de Fodor que devem ser apresentados nesse texto, mas o que mais nos chama atenção é a distinção, lembrada pelo filósofo, entre os conceitos de *intencionalidade e inteligência*. Fodor dá a entender que Searle compreendeu mal o intuito de Turing ao analisar o seu artigo (mencionado por nós anteriormente): Turing não espera que o seu teste seja uma prova cabal referente à intencionalidade em máquinas, mas apenas uma demonstração de inteligência. Para compreendermos a presença de um conceito em vez do outro no texto de Turing, basta nos voltarmos à pergunta principal da intencionalidade, quando se volta à temática da IA: o que é necessário para que um estado físico possua propriedades semânticas? Segundo Fodor, o Teste de Turing não responde essa questão, até porque *ele nunca esteve destinado ou se propôs a respondê-la*. Turing preocupa-se com outra questão, a saber: como é possível transitarmos de um pensamento verdadeiro para outro pensamento verdadeiro, sem que haja confusão ou perda de sentido? Como nossas inferências, deduções e conclusões são concatenadas de modo harmonioso no fim das contas? Esses questionamentos dizem respeito à inteligência, e embora Fodor aceite que intencionalidade e inteligência estejam ligados, nesse caso, a ausência de uma pergunta não afeta no desenvolvimento da outra. Em resumo, Fodor afirma que:

> Searle não está falando sobre inteligência, ele está falando sobre intencionalidade. Turing fala sobre inteligência, não sobre intencionalidade. Isso é bem diferente. A Inteligência Artificial é sobre inteligência, não sobre intencionalidade. Eu acredito que o pessoal da Inteligência Artificial está bem confuso com relação a isso[91]. (FODOR, 1995, p. 89)

[91] "Searle is not talking about intelligence, he is talking about intentionality. Turing talks about intelligence, not about intentionality. That is quite different. Artificial Intelligence is about intelligence, not about intentionality. I think people in Artificial Intelligence are very confused about this". (FODOR, 1995, p. 89)

Segundo Fodor (1996, p. 90), inteligência e intencionalidade podem ser tratadas de maneira isolada, uma discussão não necessariamente implica na outra. Mesmo que descobríssemos o que ocorre em um estado físico para que ele possua propriedades semânticas, ou como um mecanismo poderia possuir propriedades intencionais, ainda restaria a questão: o que faz com que um mecanismo seja inteligente? Ainda que soubéssemos o que é necessário para um pensamento ser verdade (que é, segundo Fodor, o tema da intencionalidade), ainda precisaríamos saber o que faz com que passemos de um pensamento verdadeiro para outro, de modo que nossos pensamentos não se tornem incoerentes, simplesmente um conjunto de enunciados verdadeiros sobre o mundo, mas que não possuem nenhuma ligação razoável entre si. Segundo Fodor, Searle está certo em assumir que o Teste de Turing não fornece uma resposta com relação à intencionalidade, justamente pelo fato do Teste de Turing não se propor a fazer isso. Aliás, não é a isso que a ciência cognitiva se propõe; ela se propõe a estudar a inteligência, preocupa-se em desenvolver uma teoria sintática da inteligência, não uma teoria da intencionalidade.

Por último, mas não menos importante, ainda podemos acrescentar mais uma característica da IA que talvez tenhamos ignorado: Fodor aponta a confusão de Searle ao acreditar que a IA se propõe a, com base em uma simulação, tornar possível a compreensão de toda a inteligência ou consciência, ou ainda um pouco além, que se essa simulação desse errado, isso significaria que toda a IA estaria fadada ao malogro. Fodor comenta: “A evidência é que nós não temos simulações de processos inteligentes. Nós não temos simulações, mas seria loucura ir atrás de ter esse tipo de simulação[92]” (1995, p. 91). Seria análogo a buscar a simulação de um vulcão ou mesmo de placas tectônicas em tamanho real apenas para provar que se compreende atividades sísmicas. A atividade teria sucesso se, com o pouco que se conseguiu simular, ela conseguisse ao menos: determinar algumas leis e performatizar vários testes em ambientes experimentais e artificiais, para que aquelas leis fossem postas à prova e para que alguns resultados pudessem ser previstos. Segundo Fodor, o erro de Searle é identificar a ciência cognitiva, em particular, a psicologia cognitiva, com a Inteligência Artificial, e achar que já que esta última falhou, as primeiras também falharam.

[92] “The evidence is that we don’t have simulations of intelligent processes. We don’t have simulations – but it was crazy to try to get simulations in the first place”. (FODOR, 1995, p. 91)

CONCLUSÃO

A posição de Fodor, apontada anteriormente, (sessão 4.2) poderia nos levar a uma dúvida cruel: seria possível que tudo aquilo que Searle veio objetando até então, não passou de um grande mal entendido? A IA não é o que ele acreditava que ela fosse? Mais uma vez, a resposta para essas perguntas não pode ser simples. Como tentamos apontar nesta sessão, a IA se transformou ao longo dos anos – ou melhor dizendo, as considerações acerca da IA se transformaram. Não é possível negar que existem textos que suportam a existência do que Searle chamou de IA Forte, como pudemos ver anteriormente (sessão 1.5). Entre o texto de Turing (1950), o de Searle (1980), o posicionamento de Fodor (1995) e os dias atuais, temos um intervalo de 64 anos. Como é de se esperar de uma ciência ou engenharia recente, a IA ainda está se definindo, encontrando o seu espaço no pano de fundo que já existia antes dela.

Searle elaborou a sua crítica (juntamente com o AQC) em um momento oportuno, de uma forma interessante o suficiente para que não fosse esquecida mesmo nos dias de hoje, gerando interesse filosófico e deixando muitas mentes intrigadas com alguns de seus posicionamentos e críticas. O argumento, se tomado de forma isolada e sem que seja levada em consideração a metamorfose pela qual o conceito de IA passou – e talvez ainda passe –, não pode ser considerado a contribuição de Searle. Por outro lado, deve ser levado em consideração não apenas o AQC, como o conjunto geral de críticas à IA e a justificativa para muitas dessas críticas: o conceito de naturalismo biológico. Se esses elementos forem levados em consideração de forma combinada, não será possível afirmar que os esforços de Searle devem ser desconsiderados.

Tais esforços carregam aspectos negativos e positivos sobre o problema da IA: o negativo se manifesta através das críticas, à medida que Searle aponta o que considera que *não* deveria estar ali; o positivo, como buscamos demonstrar no nosso último capítulo, se manifesta através do naturalismo biológico, como uma tentativa de propor uma nova teoria da mente – nem materialista, nem dualista. Afinal de contas, não é disso que se trata toda a discussão? O conceito de mente, de consciência. Funcionalistas, computacionalistas, dualistas de propriedades, behavioristas: a discussão entre eles não existiria se houvesse um consenso sobre esse conceito. E dessa forma, ao passo que Searle crítica a IA, bem como algumas

teorias da mente, ele só o faz porque existe a pressuposição de que o conceito de mente seja o de que o conceito de mente seja y em vez de x, onde x é o que Searle espera. Dessa forma, é importante considerar a ligação existente entre a *crítica* do modelo computacional de mente (o AQC) ao *conceito* de mente (o naturalismo biológico), proposto por Searle.

O naturalismo biológico, tampouco o AQC, são tratados contra ou a favor de uma definição última do que deveria ser a IA, ou sobre o que deveríamos esperar dos próximos computadores ou manifestações artificiais de inteligência. Para que algo assim pudesse ao menos ser sonhado, precisaríamos, antes de tudo, possuir uma definição clara e precisa do que é a mente, sem que restasse mais nenhuma dúvida. E nesse caso, talvez muito mais do que a filosofia, uma resposta seria esperada de áreas como a medicina e a neurobiologia. Disso, Searle não discorda, por isso afirma que:

> A questão sobre quais sistemas são causalmente capazes de produzir consciência e intencionalidade é uma questão factual empírica, que não será resolvida por uma teorização a priori. Já que nós não sabemos exatamente como os cérebros fazem isso, estamos em uma má posição para desvendar como outros tipos de sistemas, naturais ou artificiais, podem fazê-lo[93]. (SEARLE, 1994, p. 547)

O filósofo também concorda que não há barreiras para a consciência e a intencionalidade (SEARLE, 1994, p. 547), o que nos faz acreditar que a conclusão sobre a sua contribuição para a filosofia da mente não será, de forma alguma, um indicativo de que o debate em torno desse tema, especificamente, envolvendo a IA, irá se extinguir.

[93] "The question of which systems are causally capable of producing consciousness and intentionality is an empirical factual issue, not to be settled by a priori theorizing. Since we do not know exactly how brains do it we are in a poor position to figure out how other sorts of systems, natural or artificial might do it". (SEARLE, 1994, p. 547)

REFERÊNCIAS

ADAM, Alison. Cyborgs in the chinese room: boundaries transgressed and boundaries blurred. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

ALEKSANDER, Igor. Neural depictions of 'world' and 'self': bringing computation understanding to the chinese room. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

BISHOP, Mark. Dancing with pixies: strong artificial intelligence and panpsychism. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

BITTENCOURT, Guilherme. *Inteligência artificial*: ferramentas e teorias. 3. ed. Florianópolis: Ed. da UFSC, 2006.

BLOCK, N. J. Searle's argument against cognitive science. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

_____. Troubles with functionalism. In: _____. *Minnesota studies in the philosophy of science*. Vol. 9. Ed. C. 1978.

BRINGSJORD, Selmer; NOEL, Ron. Real robots and missing thought-experiment in the chinese room dialetic. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

CHURCHLAND, P. M.; CHURCHLAND, P. S. *Could a machine think?* Scientific American, 1990. p. 26-31

COPELAND, Jack. *Artificial intelligence*: a philosophical introduction. Blackwell Publishing, 1993.

_____. The chinese room from a logical point of view. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

_____. (Org.) *The essential Turing*: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of enigma. Oxford: Oxford University Press, 2004.

COULTER, Jeff; SHARROCK, Wes. The hinterland of the chinese room. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

CUMMINS, R. *Programs in the explanation of behavior*. Philosophy of Science, 1977.

DREYFUS, Hubert. *What computers still can't do*. New York: MIT Press, 1979.

EISENSTADT, S. A.; SIMON, H. A. A chinese room that understands. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

FREY, P. W. An introduction to computer chess. In: _____. *Chess skill in man and machine*. Ed. P. W. Frey. New York, Heidelberg, Berlin: Springer Verlag, 1977.

HARNAD, Stevan. Minds, machines, and Searle 2: what's right and wrong about the chinese room argument. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

HAUGELAND, John. Syntax, semantics, physics. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

HAUSER, Larry. Nixin' goes to China. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

HILBERT, David. Mathematical problems. In: _____. *Bulletin of the american mathematical society*. Vol. 8. No. 10. 1902.

McCULLOCH, W.S. e PITTS, W.H. A logical calculus of the ideas immanent in nervous activity. In: _____. *Bulletin of mathematical biophysics*. 1990. p.33-115.

NAGEL, T. *What is it like to be a bat?* Philosophical Review, 1974.

PENROSE, Roger. Consciousness, computation, and the chinese room. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

POPPER, K. R.; ECCLES, J. C. *The self and its brain*. Heidelberg: Springer-Verlag, 1977.

PUTNAM, H. Minds and machines. In: _____. *Dimensions of mind*. New York: Ed. S. Hook, 1960. p. 64-138.

_____. The nature of mental states. In: _____. *Art, Mind, and Religion*. Pittsburgh, University of Pittsburgh Press, 1967.

_____. The meaning of "meaning". In: _____. *Mind, language and reality*. Cambridge: Cambridge University Press, 1975.

_____. The nature of mental states. In: _____. *Mind, language and reality*.

Cambridge: Cambrigde University Press, 1975.

REY, Georges. Searle's misunderstandings of functionalism and strom AI. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

SCHANK, R. C. & ABELSON, R. P. *Scripts, plans, goals, and understanding.* Lawrence Erlbaum Press: Hillsdale, 1977.

SEARLE, J. R. In: _____. *A companion to the philosophy of mind*. Edited by Samuel Guttenplan. Cambridge: Blackwell Publishers Inc., 1996. p. 544-550.

_____. *A redescoberta da mente*. São Paulo: Martins Fontes, 2006.

_____. *Intencionalidade*. São Paulo: Martins Fontes, 2007.

_____. *Is the Brain's Mind a Computer Program?* Scientific American, 1990.

_____. *Mind*: a brief introduction. Oxford: Oxford University Press, 2004.

_____. *Minds, brains and programs.* Behaviorial and Brain Sciences, 1980.

_____. *Minds, Brains and Science*. London: Penguin, 1989.

_____. *The Mistery of Consciousness*. New York: New York Review of Books, 1997.

_____. Twenty-one years in the chinese room. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

TURING, A. M. *Computing Machinery and Intelligence*. Mind, 1950.

_____. On computable numbers, with an application to the Entscheidungsproblem. In: _____. *Proceedings of the London Mathematical Society*. Series 2. Vol. 42. 1936. p. 65-230.

WARWICK, Kevin. Alien encounters. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

WHEELER, Michael. Change in the rules: computers, dynamicam systems, and Searle. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.

WINOGRAD, Terry. Understanding, orientations, and objectivity. In: _____. *Views into the chinese room*: new essays on Searle and artificial intelligence. Edited by John Preston and Mark Bishop. New York: Oxford University Press, 2007.