UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

JOÃO VITER MARTINS TEIXEIRA

CÉREBRO E BITS:
LIMITES CONCEITUAIS DA INTELIGÊNCIA ARTIFICIAL

Rio de Janeiro

2021

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

JOÃO VITER MARTINS TEIXEIRA

CÉREBRO E BITS:
LIMITES CONCEITUAIS DA INTELIGÊNCIA ARTIFICIAL

Trabalho de Conclusão de Curso apresentado à banca examinadora do Intituto de Filosofia e Ciências Sociais da Universidade Federal do Rio de Janeiro (UFRJ) como requisito para obtenção do título de Licenciatura em Filosofia.

Orientador: Prof. Dr. Rodrigo Azevedo dos Santos Gouvea

Rio de Janeiro
2021

CÉREBRO E BITS:

LIMITES CONCEITUAIS DA INTELIGÊNCIA ARTIFICIAL

João Viter Martins Teixeira
Orientador: Professor Doutor Rodrigo Azevedo dos Santos Gouvea

Trabalho de Conclusão de Curso submetido ao Instituto de Filosofia e Ciências Sociais da Universidade Federal do Rio de Janeiro – UFRJ, como parte dos requisitos necessários para a obtenção do título de Licenciatura em Filosofia.

Nota atribuída: 10.0

Examinado por:

Professor Doutor Rodrigo Azevedo dos Santos Gouvea
Instituto de Filosofia e Ciências Sociais (UFRJ)

Professor Doutor Wilson John Pessoa Mendonça
Instituto de Filosofia e Ciências Sociais (UFRJ)

Professor Doutor Paulo Mendes Taddei
Instituto de Psicologia (UFRJ)

Outubro de 2021

AGRADECIMENTOS

A meus pais, ao meu irmão e a toda minha família, agradeço o apoio durante minha formação.

Ao Professor Rodrigo Gouvea, que gentilmente aceitou me orientar, agradeço pela paciência e seriedade com a qual me auxiliou na construção deste trabalho.

Agradeço aos professores e professoras do Instituto de Filosofia e Ciências Sociais da Universidade Federal do Rio de Janeiro, que durante todo o curso contribuíram com competência para a minha formação acadêmica.

Aos Professores Wilson John Pessoa Mendonça e Paulo Mendes Taddei, que aceitaram o convite para participarem da banca examinadora deste trabalho, agradeço pelas sugestões apresentadas.

RESUMO

O objetivo do presente trabalho é discutir possíveis respostas à pergunta: “Máquinas podem pensar?”. As perspectivas de Alan Turing e do chamado Funcionalismo de Máquinas estabeleceram um cenário teórico que apontaria para uma resposta positiva à pergunta supracitada. Posteriormente, John Searle, através do conhecido Argumento do Quarto Chinês, apresentou obstáculos conceituais significativos para a concepção de uma Inteligência Artificial Forte. A partir desses referenciais teóricos, enfatiza-se a necessidade de que conceitos como o de mente sejam desenvolvidos de forma a contemplar uma visão interdisciplinar do tema inteligência artificial. Desse modo, a suposta capacidade desses artefatos pensarem poderá ou não lhes ser atribuída de forma mais fundamentada.

SUMÁRIO

1 INTRODUÇÃO

Da *psykhé* aristotélica ao argumento do *cogito,* as reflexões sobre a natureza de nossa mente possuem uma trajetória milenar na história da filosofia. Ainda hoje, podemos afirmar que esse é um tema que gera debates a partir dos mais diversos pontos de vista. Entretanto, com o aumento da influência dos discursos científicos em nossa sociedade, as relações envolvidas na formação e constituição da nossa mente vêm sendo cada vez mais abordadas sob a ótica da razão científica contemporânea.

Desde meados do século XX, muitos autores em filosofia se dedicaram e se dedicam à tarefa de especular sobre a mente de Deus, dos anjos e de outras coisas. O sucesso no desenvolvimento de dispositivos artificiais capazes de realizar diversas atividades que antes só seriam possíveis para nós, seres humanos, resultou, segundo Teixeira (2000), numa chamada "*tecnologia do mental"* que proporcionou a aproximação entre áreas como a ciência da computação e a psicologia. A replicação tecnológica de atitudes inteligentes por meios artificiais teve grandes consequências no debate acerca das relações entre as nossas mentes e o mundo, levando alguns a sustentarem posições como as de que *computadores digitais* devidamente programados seriam capazes de pensar como nós.

O aumento exponencial das capacidades desses dispositivos dotados de inteligência artificial, fez com que posições como a supracitada reverberassem através de alguns discursos. Nesse contexto, proponho realizar no decorrer desse trabalho uma reflexão acerca da questão "máquinas podem pensar?", visando expor alguns dos principais trabalhos filosóficos que contribuíram para o desenvolvimento desse debate.

Para tanto, partiremos do artigo "*Computing Machinery and Intelligence*", de Alan Turing, para a teoria do Funcionalismo de Máquinas e posteriormente apresentaremos a perspectiva de um dos principais críticos da afirmação de que *computadores podem pensar*, John Searle. Com isso, esperamos ser possível apresentar o desenvolvimento histórico dessa questão da filosofia da mente contemporânea que, apesar de permanecer sem um consenso claramente estabelecido, deve ser abordada na busca de "*uma terceira margem do rio ou uma perspectiva da qual possamos, quando falamos de mentes e de cérebros, distinguir entre cavaleiros e moinhos de vento*" (TEIXEIRA, 2000, p. 13).

2 O JOGO DA IMITAÇÃO

Ao longo do artigo “*Computing Machinery and Intelligence*”, publicado em 1950 na revista de filosofia *Mind*, o autor Alan Turing desenvolveu a seguinte questão: máquinas podem pensar? Considerado por muitos o “Pai da Computação”, Turing já vislumbrava frente às capacidades dos computadores de sua época se haveria a possibilidade de se produzir uma máquina capaz de pensar como nós. Como sugestão para solucionar essa questão, o matemático propõe em seu artigo um experimento que chama de “jogo da imitação”, descrito por Turing da seguinte forma:

> It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A (TURING, 1950, p. 433).

Turing ainda acrescenta a seguinte regra ao jogo: o jogador X deverá tentar enganar o interrogador, ao passo que o jogador Y deve tentar ajudá-lo a acertar os respectivos sexos. Ora, mas pode-se perguntar: como esse jogo de adivinhação pode auxiliar na elucidação da pergunta central do artigo “máquinas podem pensar?”. Segundo Turing:

> We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think? (TURING, 1950, p. 434).

A sugestão, portanto, é de que o questionamento acerca da possibilidade de máquinas pensarem ou não possa (e deva) ser substituído pela seguinte pergunta: uma máquina é capaz de jogar e vencer no jogo da imitação? Caso a resposta seja positiva e um computador for capaz de se passar por um humano e enganar o interrogador no jogo da imitação, então, para o matemático, a resposta da pergunta “máquinas podem pensar?” seria positiva.

Para justificar a substituição da pergunta pelo jogo da imitação, uma definição importante precisa ser estabelecida: o que o autor quer significar pela palavra “máquina” quando se pergunta, por exemplo, se uma máquina pode jogar e vencer

no jogo da imitação. Essa reflexão será feita na terceira seção do artigo, na qual Turing limita as máquinas permitidas no experimento a um tipo específico: os *computadores digitais*. Vale ressaltar que, com essa escolha de definição, Turing não direciona o debate para a hipótese de que todos os computadores digitais sejam capazes de vencer no jogo da imitação, “mas se existem computadores imagináveis que se sairiam bem” (TURING, 1950, p. 436).

2.1 MÁQUINAS DE TURING

A abordagem de Turing é fundamentalmente teórica, pois, apesar de já existirem computadores digitais na época, o modelo que será descrito por Turing é ideal, ficando conhecido desde então como “*Máquina de Turing*”. Como computadores digitais, as máquinas de Turing são dotadas de três partes essenciais: (I) uma memória capaz de armazenar informações, (II) uma unidade de execução e (III) uma unidade de controle. Com o propósito de tornar as definições de cada parte mais inteligíveis, pode-se pensar cada uma a partir de analogias com um “sistema humano”. Afinal, a ideia proposta por Turing é de que essas máquinas podem executar qualquer operação mental que um ser humano for capaz de realizar. Nesse sentido, afirma o matemático no início da quarta seção do artigo:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job (TURING, 1950, p. 436).

Podemos compreender, portanto, a *memória do computador* como sendo equivalente a papéis nos quais alguém poderia realizar diferentes operações com símbolos; a *unidade de execução* seria o que, no caso do computador digital, tornaria possível realizar fisicamente essas operações; e a *unidade de controle*, a parte responsável por garantir que nada será realizado no papel que não esteja instruído no livro de regras.

O modelo ideal das máquinas de Turing possui ainda uma importante particularidade: um espaço de memória infinito dividido em compartimentos de maneira linear, de forma que cada unidade contendo uma informação pode ser

acessada de cada vez pela máquina. Constata-se ainda que o “livro de regras” que descreve cada operação a ser realizada pelo computador é uma abstração no caso do “computador humano”. Isso porque, em condições normais, os seres humanos simplesmente se lembram do que devem fazer quando, por exemplo, respondem a uma pergunta ou resolvem uma equação[1].

A própria ideia de um “livro de regras” a ser seguido esclarece muito bem o que o matemático entende como programação, sendo o processo de descrição das etapas que devem ser seguidas por algum sistema para atingir um objetivo X ou Y, conforme dirá o autor: “To ‘programme a machine to carry out the operation A’ means to put the appropriate instruction table into the machine so that it will do A.” (TURING, 1950, p. 438). Na citação, podemos dizer que *instruction table* significa o mesmo que livro de regras no nosso contexto.

Desse modo, pode-se afirmar que a ideia de Turing é de que as máquinas descritas, os computadores digitais, quando bem programadas e em condições adequadas (com espaço de memória suficiente) podem tomar parte satisfatoriamente no jogo da imitação. Essas máquinas são definidas por operarem de um estado X bem determinado para um outro Y, estados estabelecidos por *inputs* e conectados causalmente, de forma tal que o estado Y dependerá de qual é o estado X e vice-versa.

Para explicitar melhor essa definição, suponhamos, por exemplo, uma máquina de vendas automática que disponibilize garrafas d’água por um real e latas de refrigerante por dois reais. Essa máquina aceita como *inputs* a inserção de moedas de um real e de notas de dois reais e, para tornar o exemplo mais simples, imaginemos que essa máquina não emite troco, aceitando, portanto, o *input* máximo de dois reais. Dessa forma, podemos afirmar que essa máquina possui três estados possíveis: um estado X1 inicial, no qual a máquina se encontra apenas ligada normalmente, sem nenhuma possibilidade de *output*; um estado X2 caracterizado pela possibilidade de se comprar apenas uma garrafa d’água; e um estado X3 definido pela possibilidade de se adquirir um refrigerante ou duas garrafas d’água.

[1] Talvez possamos dizer que até mesmo a nossa lembrança seria uma espécie de “consulta inconsciente” a um de livro de regras mental. Nesse sentido, ressaltar o caráter abstrato do livro de regras para os seres humanos visa apenas expressar que não analisamos conscientemente um algoritmo específico para cada atividade que realizamos.

Note que esses estados são sempre definidos pelo *input* recebido pelo sistema e pelo estado anterior no qual a máquina se encontrava. Caso essa máquina esteja em um estado X3, causado pelo *input* que consiste na inserção de dois reais, e alguém pressione o botão indicando que deseja comprar um refrigerante, a máquina irá dispensar uma lata de refrigerante como *output* dessa operação e retornará para o estado inicial X1. Já se a máquina estiver no mesmo estado X3 e alguém pressionar o botão indicando a garrafa d'água, ela irá dispensar uma garrafa d'água e passará para um estado X2, oferecendo a possibilidade de retornar como *output* mais uma garrafa d'água antes de retornar para o estado inicial X1.

O funcionamento da máquina apresentada no exemplo é análogo à forma como operam também as máquinas de Turing. Os estados internos do sistema são bem definidos pelos *inputs* recebidos (dinheiro inserido e botões pressionados), condicionando os *outputs* possíveis (dispensa de latas de refrigerante ou garrafas d'água). Portanto, o conjunto dos estados internos, *inputs* e *outputs* e a as relações causais estabelecidas entre esses elementos vão caracterizar a máquina em questão.

Sendo assim, pelo que já foi exposto até então, podemos afirmar que o artigo de Turing situa duas questões centrais. A primeira é se, de fato, os computadores digitais podem jogar e vencer no jogo da imitação, enganando o ser humano; e a segunda é se o teste proposto por Turing é suficiente para concluir que as máquinas em questão apresentam algo que possa ser chamado de inteligência ou pensamento. Acerca da primeira, Turing prevê:

> I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning (TURING, 1950, p. 442).

Sobre essa previsão, David Dowe e Graham Oppy (2020) afirmam que ainda pode ser uma realidade distante. Porém, isso não impede que consideremos o teste de forma mais idealizada. Desse modo, o que parece mais relevante para uma análise sobre a validade do jogo da imitação e das propostas de Turing é a segunda questão mencionada anteriormente: se o teste de Turing é suficiente para que possamos atribuir inteligência às máquinas.

Para desenvolver uma reflexão sobre a questão supracitada, iremos explorá-la a partir de respostas feitas pelo próprio Turing em seu artigo a objeções que poderia

receber contra o seu teste. Dentre as objeções consideradas por Turing, aquelas que parecem ter mais relevância para nós são: a 4a objeção, intitulada “Argument from Consciousness”; a 5a objeção, intitulada “Arguments from Various Disabilities”; a 6a objeção, intitulada “Lady Lovelace's Objection”; e a 8a objeção chamada de “The Argument from Informality of Behaviour”. Entretanto, em linhas gerais, os argumentos presentes na 8a e na 5a objeção são muito semelhantes aos da 4a e da 6a objeção, de forma que, para evitar redundância sem deixar de contemplar as críticas mais fortes ao artigo de Turing, iremos discutir em mais detalhes apenas as objeções “Argument from Consciousness” e “Lady Lovelace's Objection”.

2.2 ARGUMENTO DE CONSCIÊNCIA

Turing inicia sua resposta ao argumento da consciência com uma citação de Geoffrey Jefferson que, segundo Turing, expressa bem uma crítica ao seu artigo:

> 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. 445 - 446).

Esse argumento sustenta, portanto, uma diversidade de ideias que podem ser abordadas separadamente da seguinte forma: (I) a única forma de saber se uma máquina de fato pensa seria ser aquela máquina, ou seja, apela-se ao caráter fenomenal da experiência do pensamento; (II) para que se afirme a existência de algo que possamos chamar de mente há uma necessidade de que não apenas se pense, mas que se saiba que está pensando quando o faz (*“not only write it but know that it had written it”);* (III) apresenta-se a impossibilidade de se conceber uma inteligência desvinculada dos nossos desejos e emoções, que desempenham um papel fundamental na construção do que chamamos de mente.

À primeira ideia, Turing oferece uma resposta segundo a qual nós, ao aceitar o apelo ao caráter fenomenal da experiência do pensamento, acabaríamos por aceitar também um ponto de vista solipsista. Isso porque não teríamos mais razões para acreditar que uma máquina pensa do que para acreditar que outra pessoa também

pensa, como dirá Turing: "*according to this view the only way to know that a man thinks is to be that particular man"* (TURING, 1950). Portanto, para evitar que isso ocorra, seria preciso rejeitar a ideia como um todo.

Às outras duas, Turing responde supondo um diálogo entre uma pessoa e uma máquina de escrever sonetos[2]. Nesse diálogo, a pessoa pergunta à máquina sobre o soneto que ela escreveu e a máquina é capaz de dar respostas que seriam normalmente interpretadas pelo interrogador como evidências de que aquele soneto foi escrito por uma pessoa que sente e que escreve a partir de suas memórias e de sua vida. Assim, essa situação análoga ao jogo da imitação seria, para Turing, suficiente para provar que temos tantas razões para supor que uma máquina sente algo, quanto para supor que outra pessoa sente algo.

2.3 OBJEÇÃO DE LADY LOVELACE

A objeção de Lady Lovelace à qual Turing se refere diz, em linhas gerais, que uma máquina não possui capacidade de produzir nada de novo, pois não faz nada que nós não a ordenamos a fazer. Na citação de Lady Lovelace feita por Turing (1950): "*[a machine] has no pretensions to originate anything. It can do whatever we know how to order it to perform*". Nesse sentido, o teste de Turing não seria suficiente para atribuir pensamento a uma máquina, na medida em que essa não seria capaz de produzir algo original como nós. Contra essa objeção, Turing responde citando a frase bíblica: *"There is nothing new under the sun"*.

Desenvolvendo em mais detalhes o que Turing busca expressar pela frase citada, irei me ater à consideração acerca da máquina ser capaz ou não de originalidade, tomando o termo a partir da definição apresentada por David Dowe e Graham Oppy:

> (...) in the relevant sense of *origination*, human beings "originate something" on more or less every occasion in which they engage in conversation: they produce new sentences of natural language that it is appropriate for them to produce in the circumstances in which they find themselves (GRAHAM; DOWE, 2020).

[2] É interessante pontuar que, hoje, já existem máquinas capazes de compor músicas a partir de um banco de dados e também de escrever sonetos.

Segundo essa definição de originalidade, o teste de Turing seria adequado para julgar se um computador digital de fato é capaz de produzir algo de novo. Nesse sentido, tanto Lady Lovelace estaria certa quando diz que uma máquina "*can do whatever we know how to order it to perform*", quanto Turing ao responder *"There is nothing new under the sun"*. De fato, máquinas podem fazer apenas o que as ordenamos a fazer. Dessa forma, se formos capazes de ordenar que ela execute um processo semelhante ao que nós realizamos quando respondemos a uma pergunta ou escrevemos um poema, por exemplo, seríamos capazes de dar também à máquina capacidade de criar algo novo como nós.

O que entra em questão é o que significa produzir algo original, e sobre isso dirá Turing numa curta passagem: *"Who can be certain that "original work" that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles" (*TURING, 1950). Sendo o caso que nossa originalidade, na verdade, parte de princípios gerais e aprendizado, computadores digitais seriam capazes de criar algo novo na medida em que seguissem esses princípios e pudessem aprender novas formas de agir.

Outra formulação que Turing responderá desse mesmo problema é a de que máquinas não conseguem nos surpreender. O matemático dirá que frequentemente se vê surpreendido por computadores, ora porque calcula mal, ora com pressa ou assumindo incertezas, recebendo resultados inesperados. Porém, tanto essa formulação colocada por Turing, quanto a sua resposta, parecem desviar o foco central da objeção de Lady Lovelace, isso porque, ainda que de fato esse tipo de surpresa possa ocorrer, disso não se segue que o computador produz algo de original.

Tendo respondido às objeções, Turing encerra o seu artigo expressando um certo otimismo em relação ao seu teste e ao fato de que ele pode ser suficiente para atribuir pensamento às máquinas. Para o autor, é apenas uma questão de tempo para que os avanços nas áreas da programação e da engenharia consigam dar conta de colocar o teste em prática. Porém, é válido ressaltar que Turing não oferece um argumento verdadeiramente forte para um modelo computacional da mente ou para afirmar que computadores podem, de fato, pensar. Como o próprio Turing reconhece no início da última seção de seu texto: "I have no very convincing arguments of a positive nature to support my views" (TURING, 1950, p. 454).

2.4 CONSEQUÊNCIAS DO JOGO DA IMITAÇÃO

Ainda que a visão de Turing possa ser interpretada como uma perspectiva que reduz a mente ao comportamento[3], o seu teste reverberou ao longo do tempo e suas ideias foram de grande influência no desenvolvimento da filosofia da mente durante o século XX. Apesar de suas limitações, como dirá Andy Clark: "*Turing's work clearly suggested the notion of a physical machine whose syntax-following properties would enable it to solve any well-specified problem*" (CLARK, 2001, p. 11). A definição de computação que Turing desenvolve[4] e o vislumbre de uma máquina capaz de resolver problemas através de operações sintáticas são aspectos muito relevantes para considerações posteriores sobre a possibilidade de atribuir pensamento às máquinas.

O artigo de Turing (1950) deu as bases para que o que há de essencial acerca do mental fosse atribuído não mais a estados materiais específicos do cérebro ou dos neurônios, mas sim a uma estrutura abstrata de funcionamento de nossas mentes que poderia ser replicada por computadores digitais. Nesse sentido, podemos dizer que Turing influenciou uma importante teoria computacional da mente conhecida como funcionalismo de máquinas, a qual analisaremos no capítulo seguinte.

[3] Ver o capítulo 3 de *Philosophy in a New Century: Selected Essays* (SEARLE, 2008).
[4] Como manipulação de símbolos a partir de um conjunto de regras específico.

3 O FUNCIONALISMO DE MÁQUINAS

Como vimos, o desenvolvimento do artigo de Turing não nos oferece argumentos fortes acerca da possibilidade de máquinas serem capazes de pensar. Para tanto, seria necessária uma análise que englobasse não apenas uma proposta de teste, mas uma reflexão sobre a própria natureza dos nossos estados mentais, natureza essa que poderia, ou não, ser realizada por máquinas.

Apesar disso, a sugestão de Turing de que a pergunta “máquinas podem pensar?” poderia ser substituída pela pergunta “existe algum modelo ideal de computador digital capaz de enganar um ser humano durante um teste de Turing?”, já parece se comprometer com uma visão específica acerca dos nossos pensamentos. Segundo essa perspectiva, nossos estados mentais poderiam ser descritos como estados específicos conectados entre si e responsáveis por produzir algum tipo de *output* (como seriam as respostas geradas pelo computador no caso dos testes de Turing, por exemplo).

Essa concepção inspirou uma corrente filosófica que ficará conhecida como funcionalismo.[5] De acordo com a visão funcionalista, os estados mentais (dor, medo, fome, crenças, etc) são descritos pelas funções que exercem nos sistemas em que ocorrem. Isso significa que quando sentimos dor, por exemplo, a dor que sentimos não é descrita como um estado específico do cérebro ou pelo nosso comportamento de gritar e chorar, mas antes como um estado do nosso organismo que é determinado pela função que ele irá exercer no sistema em que está inserido.

É importante ressaltar que o funcionalismo representa uma família de teorias que possui três fontes principais distintas, como dirá Ned Block:

> Functionalism has three distinct modern-day sources. First, Putnam and Fodor saw mental states in terms of an empirical computational theory of the mind. Second, Smart’s ‘‘topic-neutral’’ analyses led Armstrong and Lewis to a functionalist analysis of mental concepts. Third, Wittgenstein’s idea of meaning as use led to a version of functionalism as a theory of meaning, further developed by Sellars and later Harman. (BLOCK, 2007)

[5] A defesa feita por Turing de que as máquinas poderiam replicar atividades mentais humanas já pode ser apontada como uma ideia motivadora do aparecimento das teorias funcionalistas, como sugere Teixeira (2000, p. 125), “A noção de uma inteligência artificial como realização de tarefas inteligentes - ou seja, a possibilidade de replicação mecânica de segmentos da atividade mental humana - por dispositivos que não têm a mesma arquitetura nem a mesma composição biológica e físico-química do cérebro foi a grande motivação para o aparecimento das teorias funcionalistas”.

No presente texto, teremos como base a versão de funcionalismo sustentada por Putnam (funcionalismo de máquinas), uma vez que se relaciona melhor com a discussão que nos propomos a realizar. Sendo assim, para compreender essa perspectiva, devemos partir da consideração de que qualquer ser capaz de pensar, sentir, etc, é o que se chama de um autômato probabilístico.

3.1 AUTÔMATOS PROBABILÍSTICOS

Para esclarecer o que pode ser compreendido como um autômato probabilístico, retomemos a noção de máquina de Turing: uma máquina de Turing opera através de transições ordenadas entre estados que, como já foi apresentado anteriormente, são determinados por suas relações causais entre si e entre o conjunto dos *inputs* e *outputs* possíveis para aquela máquina. Normalmente, esses estados oscilam de forma bem determinada por esses fatores, ou seja, quando está em um estado X, por exemplo, o que dirá se o sistema irá para um próximo estado Y normalmente é uma condição estabelecida pela tabela de instruções dessa máquina e pelos *inputs*.

Sobre os autômatos probabilísticos, cito Putnam: "*The notion of a Probabilistic Automaton is defined similarly to a Turing Machine, except that the transitions between 'states' are allowed to be with various probabilities rather than being 'deterministic'*" (PUTNAM, 1975, p. 433). Essas diversas probabilidades que a tabela de instruções transmite no caso dos autômatos probabilísticos é, portanto, a principal diferença entre esses sistemas e as máquinas de Turing. Podemos compreender essa diferença da seguinte forma: suponhamos que uma pessoa receba um *input* sensorial; esse *input* ao invés de determinar uma única reação motora (ou um *output* motor), irá estabelecer as possíveis reações desse sistema a esse mesmo *input* sensorial recebido.

Dessa forma, os autômatos probabilísticos podem ser definidos pelo conjunto de seus estados possíveis junto às relações probabilísticas que orientam as transições entre esses estados e seus *inputs* e *outputs.* É dessa perspectiva que Putnam especifica uma noção que chamará de "Descrição" desses sistemas: "*A Description of S where S is a system, is any true statement to the effect that S possesses distinct states S1, S2 ... Sn which are related to one another and to the*

motor outputs and sensory inputs by the transition probabilities given in such-and-such a Machine Table" (PUTNAM, 1975).

A "*Machine Table*" que irá determinar as probabilidades de transição entre os estados do sistema corresponde ao que se chamará de Organização Funcional. Ou seja, a Organização Funcional de um sistema é o que estabelece a forma com a qual as partes deste sistema irão se relacionar. Sendo assim, conhecendo a Organização Funcional de um autômato probabilístico qualquer se conhece o que irá ditar o estado seguinte desse autômato: um conjunto de possibilidades orientadas pela função que cada uma exercerá; uma condição não tão determinística como no caso das tabelas de instruções das máquinas de Turing.

Cada estado pelo qual esse sistema transitará vai estar inserido na Descrição desse sistema e sujeito à se relacionar com outros estados possíveis e com *inputs* e *outputs* de acordo com sua Organização Funcional. Assim, um estado S1 qualquer no qual um sistema está em um dado momento implicará sempre um tipo específico de Organização Funcional em sua própria definição, podendo por isso ser chamado de Estado Total daquele sistema no momento:

> The Machine Table mentioned in the Description will then be called the Functional Organization of S relative to that Description, and the Si such that S is in state Si at a given time will be called the Total State of S (at the time) relative to that Description (PUTNAM,1975, p. 434).

Nesse sentido, quando alguém sente dor, ou fome, ou recebe qualquer outro *input* sensorial, o estado seguinte dessa pessoa seria determinado pelo Estado Total atual do sistema e, consequentemente, pela sua Organização Funcional. Desse modo, sentimentos como a dor poderiam ser descritos a partir do papel que exercem no sistema, não recorrendo à referência a um estado físico específico do cérebro ou ao comportamento de alguém, por exemplo:

> Namely, the functional state we have in mind is the state of receiving sensory inputs which play a certain role in the Functional Organization of the organism. This role is characterized, at least partially, by the fact that the sense organs responsible for the inputs in question are organs whose function is to detect damage to the body, or dangerous extremes of temperature, pressure, etc., and by the fact that the 'inputs' themselves, whatever their physical realization, represent a condition that the organism assigns a high disvalue to (PUTNAM, 1975, p. 438).

3.2 A MÚLTIPLA REALIZAÇÃO DOS SISTEMAS FUNCIONAIS

Ao buscar estabelecer uma descrição mais detalhada acerca de qual seria a natureza dos nossos estados mentais, o funcionalismo de máquinas permite que se faça uma análise mais sólida sobre a possibilidade de que computadores possam pensar. Resumidamente, o avanço proporcionado pela perspectiva funcionalista apresentada para a elaboração de uma teoria computacional da mente está em permitir que encontremos nos chamados estados funcionais um denominador comum entre os nossos estados mentais, os estados de processamento de um computador e os de algum outro sistema qualquer.

A ideia é de que o essencial para afirmar ou não a realização de algum estado mental X ou Y por algum sistema está na organização interna e abstrata desse sistema, i.e., na descrição formal dos estados X ou Y realizados nesse sistema, de forma que o mesmo estado mental poderia ser realizado tanto em nossos cérebros como em um computador digital. Nesse sentido, cito Andy Clark:

> To be in such and such a mental state is simply to be a physical device, of whatever composition, that satisfies a specific formal description. Mindware, in humans, happens to run on a meat machine. But the very same mindware (as picked out by the web of legal state transitions) might run in some silicon device, or in the alien organic matter of a Martian (CLARK, 2001, p. 14).

O que descreveria um sistema qualquer seria o conjunto dos estados internos possíveis desse sistema e suas respectivas relações causais com os *inputs* e *outputs* fornecidos e gerados por esse dispositivo. Portanto, a realização concreta de um sistema passaria a poder variar, possibilitando que um mesmo sistema seja constituído ora por um material ora por outro completamente diferente.

Caso imaginemos uma ratoeira, por exemplo, poderíamos descrevê-la como um dispositivo que, ao receber o *input* do rato passando por ela, retorna o *output* do rato aprisionado. Nesse sentido, podemos chamar de ratoeira todo sistema que apresenta essa mesma conclusão *input-output*, independente de qual seja o material que o constitui: o que define a ratoeira é a sua *funcionalidade.*

O exemplo da ratoeira talvez seja simples demais, porém, destaca bem uma ideia central: a múltipla realização dos sistemas funcionais, ou seja, o fato de que um mesmo conjunto de estados internos, *inputs* e *outputs* pode ser realizado em diferentes materiais. Essa é a noção à qual Clark faz referência quando diz que o

nosso *mindware* (conjunto dos estados internos, *inputs* e *outputs* que constituem o que chamamos de mente) ocorre em carne e osso, mas poderia ser o caso que esse sistema ocorresse “em algum dispositivo de silício, ou na matéria orgânica de um Marciano” (CLARK, 2001, p. 14, tradução nossa).

3.3 COMPUTADORES DIGITAIS PODEM PENSAR ?

Retomando a pergunta “computadores digitais podem pensar?”, inicialmente desenvolvida por Turing através da proposta do jogo da imitação, podemos concluir, portanto, que o funcionalismo de máquinas também oferece uma resposta afirmativa para essa questão. Partindo da definição de autômato probabilístico (assumidamente inspirada pelas máquinas de Turing), o centro da discussão é transferido da matéria para a forma, dando força a uma perspectiva segundo a qual o nosso cérebro poderia ser compreendido como uma “máquina de carne”:

> This notion of the brain as a meat *machine* is interesting, for it immediately invites us to focus not so much on the material (the meat) as on the machine: the way the material is organized and the kinds of operations it supports (CLARK, 2001, p. 7).

Continua Clark:

> What we confront is thus both a rejection of the idea of mind as immaterial spirit-stuff and an affirmation that mind is best studied from a kind of engineering perspective that reveals the nature of the machine that all that wet, white, gray, and sticky stuff happens to build (CLARK, 2001, p. 7).

O que o funcionalismo de máquinas busca realizar é justamente essa análise a partir de uma “*engineering perspective*”. Concebendo o cérebro como uma “máquina de carne”, busca-se refletir acerca da natureza do mental a partir da sua condição de sistema, i.e., de sua organização funcional. Como resultado dessa reflexão, observamos que, para além de uma simples analogia, ao especificar qual tipo de máquina seria o nosso cérebro, a visão funcionalista (ao menos na versão de funcionalismo que abordamos neste capítulo) afirma que operamos como autômatos probabilísticos. Dessa forma, seríamos como um “computador natural” cujas funcionalidades poderiam ser realizadas também por um computador digital.

4 O ARGUMENTO DO QUARTO CHINÊS

Até o momento apresentamos um cenário teórico no qual os estados mentais foram explicados a partir da noção de computação. Do teste de Turing ao funcionalismo de máquinas, foram elaboradas perspectivas que defendem o que chamaremos a partir deste momento de inteligências artificiais fortes. Em termos gerais, sustentar a validade de inteligências artificiais fortes é conceder que computadores programados apropriadamente são capazes de pensar como nós, conforme dirá Searle:

> (...) according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations. (SEARLE, 1980, p. 417)

Portanto, defender a possibilidade de que existam inteligências artificiais fortes é o mesmo que responder positivamente à pergunta “máquinas podem pensar?”. Por outro lado, sustentar apenas a possibilidade das chamadas inteligências artificiais fracas é atribuir aos computadores a única função de ferramentas, incapazes de experienciar qualquer coisa que se assemelhe aos nossos estados mentais: “According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool” (SEARLE, 1980, p. 417).

De agora em diante apresentaremos a perspectiva segundo a qual apenas inteligências artificiais fracas são possíveis, ou seja, de que máquinas não podem pensar. Para tanto, partiremos do experimento de pensamento do quarto chinês, no qual se pretende colocar à prova as teses sustentadas pelos defensores da inteligência artificial forte.

Suponha que uma pessoa que não compreende chinês está trancada dentro de um quarto com diversas folhas em branco e uma caneta, quando recebe uma folha com diversos ideogramas chineses. Posteriormente, essa mesma pessoa recebe uma segunda folha contendo mais ideogramas em chinês e um conjunto de regras (escritas na língua nativa dessa pessoa), indicando como relacionar os caracteres da primeira folha com os da segunda. Por último, a pessoa recebe uma terceira folha

contendo novos ideogramas chineses e também um novo conjunto de regras indicando como relacionar os novos símbolos com os anteriores e que, uma vez feitas essas relações, a pessoa deve escrever certos símbolos em resposta aos contidos nessa terceira folha e retorná-los pelo mesmo local em que recebeu as folhas.

Sem que o indivíduo trancado dentro do quarto saiba, a primeira folha é na verdade um roteiro, a segunda é uma história relacionada a esse roteiro e a terceira folha constitui um conjunto de questões sobre essa história. Além disso, ao seguir o conjunto de regras que recebeu, a pessoa no quarto produz respostas às questões feitas sobre a história que são indistintas das respostas que um falante nativo de chinês daria, ainda que o indivíduo trancado no quarto não compreenda uma palavra sequer em chinês.

O que se busca demonstrar com esse experimento de pensamento é que apenas a manipulação formal dos símbolos segundo um conjunto de regras não faz com que a pessoa dentro do quarto repentinamente saiba falar chinês. Mesmo que suas respostas sejam idênticas às de um falante nativo de chinês, o indivíduo no quarto não as compreende; para ele, elas não possuem nenhum significado: “I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing” (SEARLE, 1980, p. 418).

Searle utiliza o exemplo do quarto chinês para objetar duas alegações feitas por aqueles que defendem uma inteligência artificial em sentido forte: (i) a de que computadores são capazes de compreender e, portanto, pensar; (ii) e a de que os programas computacionais são capazes de explicar o processo pelo qual nós compreendemos as coisas. Sobre a primeira, argumenta-se que, assim como o indivíduo no quarto, o computador não compreende nada em sentido próprio, uma vez que realiza mera manipulação formal de símbolos, não produzindo significados. Acerca da segunda afirmação, Searle dirá que, na medida em que os programas forem constituídos apenas como um conjunto de operações formais, não há razões para afirmar que contribuem para explicar algo sobre o entendimento:

> 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 given to suppose that they are necessary

> conditions or even that they make a significant contribution to understanding. (SEARLE, 1980, p. 418)

4.1 CRÍTICAS ÀS INTELIGÊNCIAS ARTIFICIAIS FORTES

Retomando as considerações feitas sobre o teste de Turing, o que parece se desenvolver no exemplo do quarto chinês é uma espécie de aprofundamento do argumento da consciência respondido no artigo de Turing (1950). Busca-se demonstrar que ainda que a máquina seja capaz de fornecer um *output* indiscernível de uma resposta humana, esse *output* continua esvaziado de significado. A máquina não conhece nada, apenas opera símbolos, sendo essa operação insuficiente para a formação de entendimento. Opõe-se, assim, à uma perspectiva segundo a qual: "If you take care of the syntax, the *semantics will take care of itself*." (HAUGELAND, 1981, p. 23, apud CLARK, 2001, p. 9)

Para que se possa afirmar que máquinas não atribuem significados e que não entendem nada no mesmo sentido que nós, é necessário que se especifique como esses significados são construídos e o porquê da impossibilidade de que máquinas o construam como nós o fazemos. Aceitar que máquinas são capazes de entender algo, passaria a envolver, portanto, uma discussão sobre a formação da nossa semântica. Vale ressaltar que o que está em jogo não é um sentido metafórico de entendimento, mas sim o entendimento tomado no mesmo sentido de quando dizemos: nós entendemos as coisas. Cito Searle:

> (...)Newell and Simon (1963) write that the kind of cognition they claim for computers is exactly the same as for human beings. I like the straightforwardness of this claim, and it is the sort of claim I will be considering. (SEARLE, 1980, p. 419)

De agora em diante, refletiremos acerca dos fatores que estariam envolvidos na construção do nosso pensamento e que o tornaria irreplicável pelas máquinas. Exploraremos uma série de objeções ao argumento do quarto chinês, às quais Searle responde. Essas objeções foram feitas por defensores de uma inteligência artificial em sentido forte e, portanto, sustentarão que máquinas são capazes de pensar como nós. A primeira que abordaremos é também uma das principais: a objeção dos sistemas ("The systems reply").

4.1.1 Objeção dos sistemas

A objeção dos sistemas argumenta: ainda que o indivíduo no quarto de fato não saiba falar chinês, ele constitui apenas uma parte de um sistema que, em sua totalidade, compreende chinês. Ou seja, não podemos afirmar que a pessoa no quarto entenda chinês, mas podemos afirmar que o conjunto formado pela pessoa, pelos papéis contendo os símbolos e as regras, as canetas, a mesa e etc., entende chinês: “(...) 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. 419).

Contra essa objeção, Searle dirá que ainda que o indivíduo internalize todo o conjunto de coisas envolvido no exemplo, continuará sem compreender chinês. Isso porque a operação realizada por esse indivíduo continuará sendo uma mera manipulação simbólica desprovida de significados. Caso a história em chinês conte algo sobre gatos e cachorros, por exemplo, a pessoa empregará esses termos de forma que um falante nativo possa ser enganado, pensando que aquela pessoa de fato compreende chinês e está falando conscientemente sobre gatos e cachorros quando, na verdade, nem mesmo sabe a que aqueles símbolos fazem referência. Desse modo, a correspondência entre os *inputs* e *outputs* do indivíduo que possui o sistema do quarto chinês internalizado em sua mente e a de um falante nativo de chinês, mostra-se insuficiente para atribuir algum grau de compreensão de chinês a esse sistema.

É interessante notar que, segundo esse raciocínio, entra em xeque a própria validade do teste de Turing como critério para atribuir ou não inteligência (em sentido forte) às máquinas, como reconhece o autor:

> The only motivation for saying there must be a subsystem in me that understands Chinese is that I have a program and I can pass the Turing test; I can fool native Chinese speakers. But precisely one of the points at issue is the adequacy of the Turing test. (SEARLE, 1980, p. 419)

Além disso, Searle dirá que caso aceitássemos a objeção dos sistemas como verdadeira, as consequências seriam absurdas. Isso porque se algo pudesse ser dito como dotado de inteligência apenas pelo fato de possuir um *input,* um *output* e um tipo de programa que os relacione, então a inteligência passaria a ser um conceito extremamente banalizado, aplicável para os mais diversos objetos:

> (...) if we accept the systems reply, then it is hard to see how we avoid saying that stomach, heart, liver, and so on, are all understanding subsystems, since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands. (SEARLE, 1980, p. 420)

Dessa constatação, segue-se que para defender a sua posição sem recair nessa banalização, aqueles que sustentam a validade de uma inteligência artificial forte precisam ser capazes de distinguir o que é um sistema propriamente mental de um que não é. Quando não fazem essa distinção, esses teóricos acabam por afirmar teses possivelmente problemáticas, como a citada por Searle de que termostatos podem ser tidos como dotados de crença, à qual ele responde:

> 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, 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. 420)

Searle afirma, portanto, que para ser capaz de ter crenças devem ser levadas em conta características próprias do ser humano, características essas que não são encontradas em termostatos ou em nossos estômagos, por exemplo.

4.1.2 Objeção dos robôs

A “robot reply” considerada por Searle supõe um computador colocado dentro de um robô que seria capaz de realizar atividades muito semelhantes às humanas e interagir com o mundo a partir delas. Poderia, por exemplo, enxergar através de uma câmera, se locomover, etc, tudo isso a partir do comando do “cérebro” (no caso, do computador). A ideia por trás dessa objeção é a de que, por esse conjunto de capacidades, poderíamos afirmar que um robô desse tipo é capaz de entender as coisas como nós.

Um primeiro aspecto que Searle nota nessa réplica é que a compreensão sobre o entendimento não se limita a uma manipulação formal de símbolos, na medida em que aponta para o fato de que entender algo genuinamente é um processo que envolve relações causais com o mundo. O reconhecimento desse fato parece

concordar com uma dimensão fundamental do que Searle chama de “intencionalidade”, uma qualidade que seria crucial para o estabelecimento da cognição. A intencionalidade corresponde a essa propriedade que diversos estados mentais teriam de ser *sobre alguma coisa* (cf. SEARLE, 1983, p. 1). Caso eu tenha uma crença, por exemplo, de que o livro ao meu lado tem uma capa branca, essa crença só ocorre na medida em que é direcionada para algo no mundo, é uma crença *sobre algo*. Sendo sobre algo, essa crença expressa uma relação causal do estado mental com o mundo e nos revela, portanto, uma *intenção*.

Entretanto, Searle argumenta que mesmo partindo dessa definição não poderíamos dizer que o robô tem alguma intencionalidade, ou que compreende alguma coisa de fato. Isso porque, ainda que o robô estabeleça alguma relação causal com o mundo, o que faz, em última instância, continua sendo apenas uma manipulação formal de símbolos. O computador que coordena as ações do robô compreende as informações que articula tanto quanto o indivíduo no quarto compreende chinês. Dessa forma, Searle conclui:

> (...) I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I [the computer] have no intentional states of the relevant type. All I [the computer] do is follow formal instructions about manipulating formal symbols (SEARLE, 1980, p. 420).

Podemos perceber a partir dessa resposta, novamente, um aspecto fundamental da crítica de Searle à inteligência artificial forte: que as relações sintáticas não dão conta da construção semântica. Há aqui uma grande diferenciação em relação à perspectiva do funcionalismo de máquinas. No funcionalismo os aspectos essenciais do mental poderiam ser replicados por máquinas na medida em que seriam associados não ao material constituinte de nosso corpo mas à “(...) inner organization, that is to say, to the golden web: to the abstract, formal organization of the system” (CLARK, 2001, p. 14); poderiam ser replicados na medida em que constituem um conjunto de regras sintáticas. Dessa forma, o que parece ser o caso para Searle é que há algo essencialmente biológico na formação do nosso pensamento.

4.1.3 Objeção do cérebro simulado

A objeção do cérebro simulado supõe um programa que replique o próprio funcionamento físico do cérebro. Quando, por exemplo, esse programa respondesse perguntas em chinês, ao invés de o fazer representando informações e as articulando segundo regras específicas, a máquina reproduziria a atividade neurofisiológica que ocorre no cérebro de um falante nativo de chinês. A ideia é a de que assim o computador poderia ser capaz de compreender de fato alguma coisa. Caso contrário, o próprio entendimento do falante nativo também entraria em questão.

Primeiramente, Searle aponta que essa objeção entra em conflito com uma ideia que é central para a defesa da inteligência artificial forte, a saber, a noção de que os aspectos neurofisiológicos não seriam necessários para a formação do pensamento. De fato, a concepção funcionalista, por exemplo, defende que a realização material específica dos nossos cérebros não é essencial para a construção dos nossos estados mentais, sendo o sistema funcional comportado por essa "máquina de carne" o que realmente interessa para a questão, como foi visto anteriormente. Sobre essa visão, retomo a citação de Clark:

> To be in such and such a mental state is simply to be a physical device, of whatever composition, that satisfies a specific formal description. Mindware, in humans, happens to run on a meat machine. But the very same mindware (as picked out by the web of legal state transitions) might run in some silicon device, or in the alien organic matter of a Martian (CLARK, 2001, p. 14).

Desse possível conflito entre a objeção e a visão que ela busca defender poderia-se concluir que, na verdade, ela estaria em acordo com o que Searle busca estabelecer no exemplo do quarto chinês. No entanto, Searle argumenta que ainda que se aceite a ideia do cérebro simulado, essa máquina continuaria incapaz de entender verdadeiramente alguma coisa.

Isso se daria porque mesmo ao simular o funcionamento dos neurônios no cérebro, a máquina não seria capaz de gerar estados intencionais (ter intencionalidade), estados entendidos por Searle da seguinte forma: "It is characteristic of [intentional states] that they either are essentially directed (...) or at least they can be directed (...)" (SEARLE, 1983, p. 2). Ou seja, ainda que seja replicada a estrutura neurofisiológica do cérebro, o resultado desse processo não terá

nenhum direcionamento para algo no mundo; os significados produzidos pelo cérebro simulado ainda são vazios.[7]

4.1.4 Objeção da combinação

A última objeção que iremos abordar é a objeção da combinação (“The combination reply”) que, como indicado em seu nome, consiste numa junção das três objeções apresentadas até o momento. Sendo assim, propõe-se como refutação do argumento do quarto chinês o seguinte: a um robô dotado de um cérebro mecânico que simula as nossas sinapses nervosas, cujo comportamento é idêntico ao comportamento humano, sendo considerado como um sistema único, seria inevitável a atribuição de intencionalidade e, portanto, de pensamento e entendimento no mesmo sentido em que nós pensamos e entendemos.

Searle aceita que, de fato, nesse caso estaríamos todos propensos a dizer que esse robô está pensando (ou sequer saberíamos diferenciar o robô de um ser humano), mas no momento em que soubéssemos que seu comportamento é resultado da execução de um programa de mera manipulação de símbolos, Searle afirma que deixariamos de crer que há ali algum pensamento. Isso se dá pois a intencionalidade aparece como um fenômeno impossível de ser realizado nesse caso, na medida em que não pode ser fruto de uma manipulação de símbolos esvaziados de significado e independente das relações causais produzidas pela matéria na qual o sistema ocorre.

4.2 A MENTE NÃO INDEPENDE DA MATÉRIA

Em suma, a ideia que paira sobre todas as respostas dadas por Searle é a de que, ao contrário do que alguns modelos de inteligências artificiais fortes defendem, a nossa mente não pode ser tomada como uma estrutura meramente formal para a qual a matéria seria irrelevante:

> It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality

[7] A justificativa de Searle parece apontar numa direção semelhante à visão de Hacker e Bennet sobre a atribuição de atributos psicológicos a partes de um animal ou a partes de seu cérebro, cito: “(...) são os seres humanos que pensam e raciocinam, e não os seus cérebros. O cérebro e as suas actividades *tornam possível* para *nós* - não para *ele* - percepcionar e pensar, sentir emoções, e formar e realizar projectos” (HACKER e BENNETT, 2003, p. 18).

> (...) it is because I am a certain sort of organism with a certain biological (i.e. chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena (SEARLE, 1980, p. 422).

Partindo dessa valorização do componente material das nossas mentes, o argumento do quarto chinês irá estabelecer uma diferença fundamental entre uma "aparência de compreensão" e uma "compreensão genuína", apontada por Sathler (2010, p. 134). Na medida em que os computadores apenas realizam manipulações sintáticas, seriam capazes de produzir uma mera compreensão aparente, operando símbolos de acordo com um conjunto de regras, mas não atribuindo significado nenhum a esses símbolos. O ser humano, ao contrário, por ser um organismo dotado de uma intencionalidade, introduz significado, ou "direção", nos símbolos, sendo capaz de compreendê-los genuinamente.

5 O QUE OS COMPUTADORES NÃO CONSEGUEM ENTENDER

Desse momento em diante, iremos estender a compreensão de Searle sobre a questão das inteligências artificiais a partir de seu livro *Philosophy in a New Century: Selected Essays*, publicado em 2008, e de seu artigo intitulado *What your computer can't know*, publicado em 2014. A ideia é que a concepção defendida no argumento do quarto chinês se torne mais clara, para que, a partir da articulação com o que foi exposto nas seções acima, possamos pensar o que significam os resultados da evolução do projeto da inteligência artificial desde Turing até a contemporaneidade, bem como as possibilidades e atribuições que são devidas ao computadores em relação ao ato de pensar e ter estados mentais.

Anos após a publicação do seu artigo "Minds, brains, and programs" (1980), Searle ainda sustenta que a tese principal defendida na apresentação do argumento do quarto chinês[8] continua inabalada. Essa tese, apresentada na seção anterior, é retomada por Searle (2008) a partir de três distinções fundamentais[9] entre sintática e semântica, entre simulação e duplicação e entre objetos que dependem e independem de um observador. A diferença entre sintática e semântica e suas implicações no desenvolvimento de inteligências artificiais foram exploradas durante a apresentação do argumento do quarto chinês e suas objeções, na seção anterior. Portanto, para que evitemos uma repetição excessiva, daremos um enfoque maior nesse primeiro momento para a distinção entre objetos que dependem e independem de um observador[10].

5.1 OBJETOS DEPENDENTES E INDEPENDENTES DE OBSERVADOR

Através da diferenciação entre esses dois conceitos, *objetos independentes de um observador* e *objetos dependentes de um observador*, Searle divide o mundo em

[8] "The fundamental claim is that the purely formal or abstract or syntactical processes of the implemented computer program could not by themselves be sufficient to guarantee the presence of mental content or semantic content of the sort that is essential to human cognition" (SEARLE, 2008, p. 67).

[9] Ver o capítulo 4 de Searle (2008).

[10] Vale ressaltar que todos os conceitos apresentados por Searle se articulam de alguma forma na construção de sua crítica. A consideração acerca das distinções citadas é feita separadamente apenas para que se torne mais clara ao leitor.

duas grandes categorias de coisas[11]: de um lado aquelas como o dinheiro, o governo, as universidades e as fronteiras entre países, que existiriam na medida em que são criadas por nossa consciência, sendo, portanto, entidades que existem de forma dependente de um observador; e de outro lado coisas como as florestas, o oceano e a força da gravidade, que existem e continuarão existindo no mundo ainda que não haja nenhum ser vivo capaz de pensar sobre elas e que, portanto, seriam independentes de um observador.

Quando pensamos, por exemplo, que um copo de café em cima da mesa está frio, para Searle trata-se de um fenômeno que, tomado em sua realidade psicológica enquanto pensamento consciente, *independe de qualquer observador*; o pensamento consciente é algo que existiria em nós, como as florestas, o oceano e a força da gravidade existem no mundo. Ao contrário, se escrevemos a frase "o copo de café em cima da mesa está frio" em uma folha de papel, a realidade do que foi escrito enquanto uma sentença com sentido só existe na medida em que é lida e interpretada por alguém. Dessa forma, a ocorrência de uma sentença com significado na folha de papel *depende de um observador*.

Essa distinção entre coisas dependentes e independentes de um observador municia uma análise de concepções centrais discutidas ao longo das seções anteriores, como *computação e inteligência.* Para que possamos situar o desenvolvimento desses debates de forma mais direcionada e recente, exploraremos uma resenha crítica realizada por Searle e publicada na revista *The New York Review of Books*, em 2014*.* Em seu artigo, Searle irá analisar dois livros a partir da distinção entre objetos dependentes e independentes de um observador, argumentando que ambos erraram em suas considerações sobre as relações entre consciência, computação, informação, cognição, e outros fenômenos[12].

O primeiro deles é *The 4th Revolution: How the Infosphere Is Reshaping Human Reality*, livro de Luciano Floridi, cuja tese principal que Searle analisará é a de que toda a realidade é, hoje, composta por um grande conjunto de informações que dá forma ao que Floridi chamará de infosfera. O outro livro analisado por Searle

[11] Searle afirma claramente a importância dessa distinção para a sua compreensão das coisas no mundo: "Absolutely essential to our understanding of the world is to be able to distinguish between those features of the world that exist independently of our attitudes and purposes, and those that exist only relative to us" (SEARLE, 2008, p. 78)

[12] Ver "*What your computer can't know*", John Searle, 2014.

ao longo do artigo é *Superintelligence: Paths, Dangers, Strategies*, de Nick Bostrom. A posição do livro de Bostrom que Searle irá criticar é a suposição de que um computador dotado de uma superinteligência[13], muito superior à humana, estaria por surgir e que seria capaz de desenvolver intenções e desejos próprios, podendo, inclusive, representar uma ameaça à raça humana.

Em relação à crítica de Searle sobre a tese de Floridi, podemos compreendê-la da seguinte forma: (i) Floridi concebe a estrutura básica de todas as coisas que existem como sendo a informação, por essa razão dá ao mundo o nome de infosfera; Searle responde que (ii) a informação não existe nas coisas de forma independente de um observador mas apenas em relação a um observador consciente. Nesse sentido, a informação não poderia ser a estrutura básica das coisas, uma vez que existe apenas na medida em que é uma "criação" da nossa consciência: "Consciousness is the basis of information; information is not the basis of consciousness" (SEARLE, 2014, p. 6).

De forma semelhante, Searle desenvolve sua resposta a Bostrom, através da qual poderemos nos direcionar para uma discussão mais específica sobre os computadores digitais. A crítica de Searle é fundamentada na distinção de duas formas de se conceber a ideia de *computação*:

> (...) it is important to see that in the literal, real, observer-independent sense in which humans compute, mechanical computers do not compute. They go through a set of transitions in electronic states that we can interpret computationally. The transitions in those electronic states are absolute or observer independent, but the computation is observer relative. The transitions in physical states are just electrical sequences unless some conscious agent can give them a computational interpretation (SEARLE, 2014, p. 2).

Portanto, enquanto atividade realizada conscientemente por nós, a computação pode ser entendida como algo que independe de um observador, como o pensamento de que há um copo de café frio em cima da mesa. No entanto, a computação operada por um computador digital possui apenas uma existência relativa a um observador e, como o pensamento presente na frase escrita "o copo de

[13] Bostrom apresenta três formas de definir o termo *superinteligência*, não obstante, reconhece uma certa equivalência entre elas. Portanto, acredito que podemos sintetizar o significado do termo pela seguinte definição dada por Bostrom: "(...) we use the term "superintelligence" to refer to intellects that greatly outperform the best current human minds across many very general cognitive domains" (BOSTROM, 2014, p. 52).

café em cima da mesa está frio", só ocorre na medida em que é interpretada por alguém.

Searle reafirma, portanto, a defesa do ponto de vista apresentado inicialmente em seu argumento do quarto chinês ao estabelecer que a computação executada por um ser humano difere quanto a natureza da computação realizada por computadores digitais. Isso porque a primeira possuiria uma realidade psicológica que a tornaria independente de qualquer observador e a última não. A própria existência de algo que possa ser chamado de computação no caso dos computadores digitais dependeria de que alguma pessoa observasse a realização dos estados físicos operados pelo computador digital e os interpretasse intencionalmente.

Assim, Searle responde a suposição de Bostrom de que um computador superinteligente poderia desenvolver vontades próprias e representar uma ameaça aos seres humanos, afirmando que é uma suposição incoerente já que no caso dos computadores não há uma inteligência que seja independente de um observador que a interprete, como a nossa:

> If we ask, "How much real, observer-independent intelligence do computers have, whether 'intelligent' or 'superintelligent'?" the answer is zero, absolutely nothing. The intelligence is entirely observer relative (SEARLE, 2014, p. 3).

5.2 DUPLICAÇÃO E SIMULAÇÃO

Podemos fazer uso da crítica de Searle a Bostrom para exemplificar uma das distinções citadas no início da sessão, entre *simulação* e *duplicação*, uma vez que a ideia é que os computadores digitais não conseguem *duplicar* o pensamento humano, já que não produzem as mesmas capacidades causais que o nosso cérebro produz (gerar uma intencionalidade consciente; atribuir significados). Mesmo que um computador possa vir a agir de forma idêntica a um ser humano, o que realiza é mera *simulação* do nosso comportamento, não uma *duplicação* dos nossos estados mentais. Duplicação difere de simulação, portanto, na medida em que a primeira trata da produção das mesmas capacidades causais em diferentes sistemas; e a segunda, ao contrário, apenas "imita" o resultado dessas capacidades causais, sendo esse resultado não produzido por essas capacidades de fato. De forma análoga, retomo a distinção de Sathler (2010) entre uma "compreensão genuína" que estaria relacionada

às capacidades causais do cérebro e uma “aparência de compreensão”, indiferente quanto às capacidades causais do cérebro, apenas simulada.

Entretanto, vale pontuar que, ao afirmar que os computadores digitais apenas simulam nossa atividade mental, Searle não estabelece uma impossibilidade de que um dia sejamos capazes de duplicar as nossas capacidades mentais por meios artificiais, mas ressalta:

> The point (...) is that any such artificial machine would have to be able to duplicate, and not merely simulate, the causal powers of the original biological machine (SEARLE, 2008, p. 72).

Continua Searle:

> an artificial brain would have to do something more than simulate consciousness, it would have to be able to produce consciousness. It would have to cause consciousness (SEARLE, 2008, p. 72).

Nesse sentido, o teste de Turing, introduzido no primeiro capítulo, jamais nos daria um parâmetro suficiente para atribuir ou não pensamento aos computadores digitais. Isso porque, mesmo vencendo no jogo da imitação, essas máquinas simplesmente não possuem o que é necessário para duplicar a mente. O computador digital nem sequer computaria no mesmo sentido em que nós computamos as coisas[14]. Para Searle, um computador digital realiza apenas operações físicas que são interpretadas por nós como computações, e qualquer atribuição para além disso dependeria de uma intencionalidade projetada por *nós* na *coisa física*: “*Computational states are not discovered within the physics, they are assigned to the physics*” (SEARLE, 2008, p. 94).

5.3 MÁQUINAS PODEM PENSAR?

Através da análise das duas críticas realizadas por Searle, podemos afirmar que um denominador comum entre elas continua sendo a noção de intencionalidade. As coisas que dependem de um observador, tal como definidas por Searle, diferem das que independem por requererem uma intencionalidade, um direcionamento por parte do observador para existirem; nós projetamos a sua existência.

[14]“(...) the characterization of a process as computational is a characterization of a physical system from outside; and the identification of the process as computational does not identify an intrinsic feature of the physics, it is essentially an observer relative characterization” (SEARLE, 2008, p. 95).

Desse cenário teórico, podemos então retomar a pergunta feita no início do primeiro capítulo: “máquinas podem pensar?”. Para Searle, a resposta é: sim, máquinas podem pensar, e máquinas de um tipo bem específico, a saber, o cérebro animal. No entanto, as ideias de Searle apontam um problema para propostas como as defendidas por Turing e pelo funcionalismo de máquinas, de que os computadores digitais seriam capazes de produzir pensamento como nós; ou que a mente seria para o cérebro como o *software* é para o *hardware*. Segundo Searle, os computadores digitais não podem pensar como nós, pois apenas realizam operações físicas desprovidas de intencionalidade, sendo a própria consideração do que fazem como um conjunto de operações sintáticas dependente de um observador intencional[15]. Ainda que os computadores sejam capazes de provocar a ilusão de estarem pensando, como no jogo da imitação de Turing, não há uma realidade naquele pensamento como há no nosso[16]. Em outras palavras, o computador só pensa na medida em que *nós projetamos o pensamento nele.* O computador digital simula a nossa atividade mental, mas não a duplica de fato. Nessa concepção reside a crítica de Searle desde o argumento do quarto chinês.

Entretanto, a visão de Searle não parece comprometer alguns aspectos introduzidos pela proposta funcionalista: Searle parece concordar que é possível replicar certos sistemas em diferentes tipos de materiais[17]. No próprio artigo “*What Your Computer Can’t Know*”, mesmo no caso do cérebro, aceita-se a possibilidade de que talvez possamos duplicar as suas funções em uma matéria completamente diferente. Nesse sentido, a crítica de Searle parece problematizar com mais enfoque a visão de que *computadores digitais seriam capazes de realizar o papel de duplicar a nossa mente*, concordando, portanto, com Teixeira nesse aspecto:

> Certamente não somos máquinas idealizadas nem tampouco nossas atividades mentais podem ser replicadas fornecendo-se delas apenas uma descrição abstrata na qualidade de um software que poderia ser rodado em qualquer tipo de máquina independentemente de sua arquitetura. Contudo, múltipla instanciação não significa instanciação irrestrita. Não é qualquer tipo de substrato físico ou de hardware que pode simular a vida mental - é isto que a neurociência tem procurado nos ensinar. Mas a neurociência não nos ensina que o cérebro é necessariamente irreplicável; tampouco que não podemos reproduzir suas características funcionais usando outros materiais e arquiteturas

[15] Ver capítulo 5 de Searle (2008).
[16] Uma realidade independente de um observador.
[17] Múltipla realização dos sistemas funcionais.

> para simular a mente - da mesma forma que uma máquina de diálise simula um rim (TEIXEIRA, 2000, p. 178).

Assim, a conclusão parece ser que enquanto não formos capazes de criar uma máquina que produza as mesmas relações causais da nossa mente, como a nossa consciência, não podemos falar em uma máquina que pense como nós. Nesse fato, o próprio Searle reconhece uma grande dificuldade[18]: entender como nosso próprio cérebro produz a consciência. Até que entendamos com mais clareza a construção da nossa própria consciência, o horizonte parece ser o de simulações cada vez mais complexas de nossa mente, mas que, apesar de qualquer nível de eficiência que essas simulações possam alcançar na solução de problemas dos mais diversos, não serão capazes de transpor os limites da imitação.

[18] “The difficulty with carrying out the project is that we do not know how human brains create consciousness and human cognitive processes” (SEARLE, 2014, p. 6)

6 CONSIDERAÇÕES FINAIS

Ao longo deste trabalho, analisamos algumas abordagens desenvolvidas em torno da questão “máquinas podem pensar?”. No primeiro capítulo, pensamos a partir do artigo “*Computing Machinery and Intelligence*” de Turing, que visava substituir a pergunta supracitada pela possibilidade, ou não, de que computadores digitais (ou Máquinas de Turing) pudessem vencer no jogo da imitação. Segundo Turing, caso os computadores tivessem sucesso nesse teste (o que Turing acreditava ser perfeitamente plausível), não teríamos um porquê para negar que pensam.

Vimos que, apesar de algumas limitações no teste de Turing, seu artigo lança as bases para o desenvolvimento de uma importante teoria da mente: o Funcionalismo de Máquinas. Exploramos essa visão a partir de Hilary Putnam e Andy Clark, apresentando a noção de autômato probabilístico, bem como a sua relação com as chamadas Máquinas de Turing. Posteriormente, demonstramos como, ao nos definir como autômatos probabilísticos, os chamados estados funcionais poderiam passar a ser um denominador comum entre os nossos estados mentais e os estados de processamento de um computador digital.

Por fim, foi explorada a perspectiva crítica de John Searle a partir do seu conhecido experimento de pensamento do quarto chinês, publicado em 1980, e de trabalhos mais recentes publicados em 2008 e em 2014. Durante essa etapa, refletiu-se sobre as razões para que computadores digitais não fossem capazes de pensar como nós, realizando apenas uma simulação de nossa atividade mental, mas não pensando de fato, tendo a *intencionalidade* como conceito central para o desenvolvimento da argumentação de Searle. Logo, concluímos que a única máquina capaz de pensar como nós, para Searle, seriam os nossos cérebros, e que qualquer outra que buscasse o mesmo *status* de ser pensante deveria não apenas *simular* a nossa mente, mas *duplicar* suas capacidades causais.

A partir dessa trajetória, podemos afirmar que a questão das inteligências artificiais ainda é um campo em pleno desenvolvimento, e que, para fundamentar algumas possibilidades e atribuições feitas em relação às máquinas “inteligentes”, não bastará apenas um trabalho filosófico-conceitual sobre ideias como mente e inteligência, mas que esse trabalho conceitual seja abordado de forma a acomodar uma perspectiva interdisciplinar com formas científicas de investigação.

REFERÊNCIAS BIBLIOGRÁFICAS

BOSTROM. Nick. **Superintelligence: Paths, Dangers, Strategies**. Oxford: Oxford University Press, 2014.

BLOCK, Ned. **Consciousness, Function, and Representation: Collected Papers, Volume 1**. Massachusetts: MIT Press, 2007.

CLARK, Andy. **Mindware: An Introduction to the Philosophy of Cognitive Science**. Estados Unidos da América: Oxford University Press, 2000.

HACKER, Peter; BENNET, Max. **Fundamentos Filosóficos da Neurociência**. Tradução: Rui Alberto Pacheco. Lisboa: Instituto Piaget, 2003.

OPPY, Graham; DOWE, David. The Turing Test. **Stanford Encyclopedia of Philosophy Archive**, n. Winter 2020, 9 abr. 2003. Disponível em: https://plato.stanford.edu/archives/win2020/entries/turing-test/. Acesso em: 9 set. 2021.

PUTNAM, Hilary. **Philosophical Papers, Volume 2.** Cambridge: Cambridge University Press, 1975.

SATHLER, André. Uma Resposta Funcionalista ao Argumento do Quarto Chinês de Searle. **Cognitio-Estudos: Revista Eletrônica de Filosofia**, São Paulo, v. 7, n. 2, p. 132-140, julho-dezembro, 2010.

SEARLE, John. Minds, brains, and programs. **The behavioral and brain sciences**, v. 3, n. 3 p. 417-424, 1980.

SEARLE, John. **Intentionality: An Essay in the Philosophy of Mind**. Cambridge: Cambridge University Press, 1983.

SEARLE, John. **Philosophy in a New Century: Selected Essays.** Cambridge: Cambridge University Press, 2008.

SEARLE, John. What your computer can’t know. **New York Review of Books**, 2014. Disponível em: https://www.nybooks.com/articles/2014/10/09/what-your-computer-cant-know/. Acesso em: 09 set. 2021.

TEIXEIRA, João de Fernandes. **Mente, cérebro e cognição**. Petrópolis, RJ: Vozes, 2000.

TURING, Alan. Computing Machinery and Intelligence. **Mind**, v. 59, n. 236, p. 433-460, outubro, 1950. Disponível em: https://academic.oup.com/mind/article/LIX/236/433/986238. Acesso em: 09 set. 2021.