UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE CIÊNCIAS HUMANAS,
COMUNICAÇÃO E ARTES

LIBNI EWERTON TELES DA SILVA

As dificuldades da Inteligência Artificial Forte: O argumento de John Searle e a teoria conexionista.

Maceió
2022

LIBNI EWERTON TELES DA SILVA

As dificuldades da Inteligência Artificial Forte: O argumento de John Searle e a teoria conexionista.

Trabalho de Conclusão de Curso da Licenciatura em Filosofia da Universidade Federal de Alagoas, como requisito parcial para obtenção do grau de licenciado em filosofia.

Orientador: Prof. Dr. Ricardo Rabenschlag

2022

Maceió

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

S586d Silva, Libni Ewerton Teles da.
As dificuldades da inteligência artificial forte : o argumento de John Searle e a teoria conexionista / Libni Ewerton Teles da Silva. – 2022.
58 f. : il.

Orientador: Ricardo Rabenschlag.
Monografia (Trabalho de Conclusão de Curso em Filosofia) – Universidade Federal de Alagoas. Instituto de Ciências Humanas, Comunicação e Artes. – Maceió, 2022.

Bibliografia: f. 57-58.

1. Searle, John R., 1932-. 2. Churchland, Paul M., 1942-. 3. Churchland, Patrícia Smith, 1943-. 3. Inteligência artificial. 4. Conexionismo. 5. Behaviorismo (Psicologia). 6. Filosofia da mente. I. Título.

CDU: 159.9.019.4

“A robot may not injure a human being or, through inaction, allow a human being to come to harm.”

I, ROBOT, Isaac Asimov

RESUMO

"Pode uma máquina pensar?". Em 1980 John Searle publicou o artigo Minds, Brains and Programs, que mudaria todo o debate em torno da chamada Inteligência Artificial Forte. Neste artigo Searle argumenta que as limitações da IA Forte delineiam-se pela sua incapacidade de geração causal de intencionalidade. Mais tarde, Searle revisitou seu argumento. Em 1990, num artigo para a revista *Scientific American*, Searle respondeu a críticas e contrapontos, refinando sua abordagem no argumento contra IA Forte. Nesse artigo, o cerne do seu argumento contra a Inteligência Artificial Forte é posto logo no subtítulo: "Mentes são semânticas, programas são sintáticos". Boa parte das réplicas mais desafiadoras contra o argumento de Searle partiam de esboços da teoria conexionista, especialmente o argumento de Paul M. Churchland e Patrícia S. Churchland, que consideraram o uso de redes neurais artificiais para o estudo da cognição humana e uma possível solução para o argumento de Searle.

Palavras-Chave: Inteligência Artificial Forte, John Searle, Conexionismo, Paul M. Churchland. Patrícia S. Churchland, Behaviorismo, Filosofia da Mente.

ABSTRACT

“Can a machine think?”. In 1980 John Searle published the article Minds, Brains and Programs, which would change the entire debate around the so-called Strong Artificial Intelligence. In this article Searle argues that the limitations of Strong AI are outlined by its inability to causally generate intentionality. Later, Searle revisited his argument. In 1990, in an article for Scientific American, Searle responded to criticisms and counterpoints by refining his approach to the argument against Strong AI. In this article, the core of his argument against Strong Artificial Intelligence is laid out in the subtitle: “Minds are semantic, programs are syntactic”. Most of the most challenging replies against Searle's argument came from sketches of connectionist theory, especially the argument of Paul M. Churchland and Patricia S. Churchland, who considered the use of artificial neural networks for the study of human cognition as a possible solution for Searle's argument.

Key-words: Strong Artificial Intelligence, John Searle, Connectionism, Paul M. Churchland. Patrícia S. Churchland, Behaviorism, Philosophy of Mind.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

RAND: Research and development Corporation

MIT: Massachusetts Institute of Technology

PDP: Parallel Distributed Processing

GPS: General Problem Solver

ADALINE: Adaptive Linear Neuron

GPT-3: Generative Pre-trained Transformer 3

SUMÁRIO

1. INTRODUÇÃO

“Máquinas pensam?” é uma pergunta ampla. Poderíamos colocar debaixo desse guarda-chuva outras questões como: “Computadores tem sentimentos?”, “Máquinas tem consciência?”, “Como saber se máquinas tem mente?”. Nesse trabalho podemos assumir que vamos lidar com praticamente todas essas perguntas, porque vamos tratar o problema de maneira geral, mas mantendo a concentração em aspectos chave da questão, de forma que a partir do que for tratado aqui, haverá possibilidades de se discutir várias questões relacionadas ao problema de máquinas com mentes. A razão disso é que apanharemos o problema de maneira mais nevrálgica, já que quando se pergunta sobre máquinas com mentes, discutimos diretamente sobre mente, Inteligência Artificial, materialismo, fisicalismo e consciência.

Não se trata de um tema simples, apesar da sua fama no senso comum, por isso vamos esclarecer aqui de antemão certas coisas antes de irmos ao que interessa. No título do trabalho temos três palavras-chave: “Inteligência Artificial Forte”, “John Searle” e “Conexionismo”. Vamos explicar a razão de elas estarem ali. A primeira coisa é que a pergunta “Máquinas pensam?” foi feita exatamente dessa forma por Alan Turing em 1950. Essa pergunta originou o que mais tarde veio a ser chamado de Inteligência Artificial e, subsequentemente, a Inteligência Artificial Forte. Também já temos diversas respostas a essa questão, nas quais a mais relevante é a que proposta pelo filósofo John Searle em 1980, que, grosso modo, se trata de um sonoro “Provavelmente, não”. Também temos várias réplicas a Searle, das quais vamos considerar apenas uma, que parte da teoria conexionista e que assume uma possibilidade real para máquinas com mentes. Fica então esclarecido o que queremos com o título: “As dificuldades da Inteligência Artificial Forte: O argumento de John Searle e a teoria conexionista”. Comecemos com Alan Turing.

Alan Turing foi um engenheiro e matemático que, entre outras coisas, trabalhou na quebra da criptografia da máquina responsável por cifrar as comunicações dos alemães durante a Segunda Guerra Mundial, a máquina se chamava ENIGMA. Posteriormente Turing contribuiu para o avanço da ciência da computação, sendo assim chamado de “O pai do computador e da inteligência artificial”. O primeiro ensaio de facto sobre Inteligência Artificial foi desenvolvido por ele, e é a partir daí onde há uma verdadeira sistematização sobre como funcionaria um programa de Inteligência Artificial. No tempo em que escreveu esse artigo, os computadores acabavam de ser tornar digitais e eram ferramentas essenciais para o trabalho militar e a pesquisa matemática complexa. Turing viu além, e sendo assim, argumentou que os computadores poderiam ser tanto máquinas que servissem para resolver problemas universais (coisa que se verifica hoje) como

poderiam pensar, e isso se provaria a partir de um teste comportamental. Falaremos mais sobre Turing e seu teste na seção seguinte.

Por definição Inteligência Artificial diz respeito a Agentes Inteligentes, que são máquinas de estados finitos capazes de perceber seu ambiente e tomar ações que maximizem suas chances de atingir seus objetivos. Uma Inteligência Artificial portanto vai além de um programa padrão que cumpre tarefas seguindo uma programação específica. Uma inteligência artificial não é explicitamente programada para fazer o que faz, mas pode sim seguir regras. Dentro da Inteligência Artificial tem-se duas terminologias, a Inteligência Artificial Fraca e a Inteligência Artificial Forte. É com a segunda que estamos preocupados, mas vamos descrever as duas. A asserção de que as máquinas talvez possam agir de maneira inteligente (ou, quem sabe, agir como se fossem inteligentes) é chamada hipótese de Inteligência Artificial Fraca pelos filósofos, e a asserção de que as máquinas que agem de maneira inteligente e estão realmente pensando (em vez de simularem o pensamento) é chamada hipótese de Inteligência Artificial Forte. Em geral, não há consenso sobre as definições, mas no meio filosófico assume-se que uma Inteligência Artificial Forte é um programa ou máquina que é senciente e consciente, sendo inclusive essa definição que John Searle vai considerar em sua crítica.

A maior parte dos pesquisadores de IA assume em princípio a hipótese de IA Fraca, e não se preocupa com a hipótese de IA forte. O termo “IA Forte”, aliás, é controverso. Alguns cientistas preferem a terminologia IA Geral, em que um agente (programa ou máquina) teria múltiplas inteligências, ou seria capaz de aprender qualquer coisa que um ser humano pode aprender. Nesse trabalho vamos utilizar o termo “Inteligência Artificial Forte” como sinônimo para “Máquinas com mentes”, isto é, máquinas capazes de pensar. Apesar de não lidarmos especificamente com a chamada IA Fraca nesse trabalho, ela também é relevante para a discussão da IA Forte, como por exemplo nos trabalhos de crítica a Inteligência Artificial do filósofo Humbert Dreyfus. Falaremos mais sobre isso no capítulo seguinte. É importante dizer que Turing não criou nenhuma dessas definições ou sequer fala sobre elas em algum lugar do seu ensaio. Essas definições são posteriores ao seu trabalho. Mesmo assim, é possível afirmar que a contribuição de Turing com seu artigo serve a Inteligência Artificial em geral, tanto a mais prática, quanto a mais hipotética. Por isso assumimos que o trabalho de Turing é o primeiro grande trabalho sobre IA Forte, mesmo que nem a própria Inteligência Artificial existisse formalmente no tempo em que ele o escreveu.

Invariavelmente, quando escreveu sobre a possibilidade de máquinas pensarem, Turing abordou uma questão filosófica, não atoa, seu trabalho foi publicado no *Mind Journal*, um dos mais

importantes periódicos sobre filosofia da mente. Podemos encontrar no texto de Turing discussões sobre questões como Behaviorismo, Dualismo, Imaterialidade da Alma, Problema das Outras Mentes, Intencionalidade, Consciência, Fisicalismo e Computacionalismo. (Algumas dessas questões foram abordadas por ele antes de existirem formalmente como objetos de estudo da filosofia da mente). Apesar dessa sopa de problemas, A Inteligência Artificial Forte é bem resolvida no que diz respeito a algumas questões. Quando falamos de IA Forte, falamos de uma posição teórica que assume o materialismo (A mente ocorre por uma ação causal e física) e poderíamos também dizer que ser favorável a IA Forte significa ser assumir o Behaviorismo, mas vamos apenas dizer que a IA Forte é bastante amparada por experimentos comportamentais, como o próprio experimento de Turing, e as vezes rejeita a ideia de que precisamos investigar além do comportamento. Turing mesmo assumia tal coisa.

Naturalmente, tanto Turing quanto a Inteligência Artificial como um todo receberam críticas. Primeiro vieram as críticas ao ensaio que Turing escreveu em 1950. Essas críticas não vamos tratar aqui (Interessados podem consultar K. Gunderson, 1964, P.H. Millar, 1973 e J. Moor, 1976). Em seguida, as críticas a Inteligência Artificial. É preciso levar em conta que o período de euforia com o desenvolvimento da Inteligência Artificial aconteceu em paralelo ao desenvolvimento de várias terias Behavioristas da Filosofia da Mente. Podemos dizer que essa euforia ocorreu entre 1956 a 1970, quando se imaginava que a Inteligência Artificial em poucas décadas alcançaria um nível humano de cognição. É difícil para qualquer entendedor da história da ciência da computação, quanto para qualquer estudante de filosofia, não traçar paralelos entre aquilo que acontecia nos laboratórios de informática e o que acontecia nos institutos de psicologia dos Estados Unidos, especialmente quando temos acesso às teorias como por exemplo Funcionalismo e a Ciência Cognitiva incipiente dos anos 1970. Nesse sentido, algumas críticas que eram feitas a essas teorias filosóficas acabavam atingindo também ao projeto da Inteligência Artificial Forte, como por exemplo, o famoso ensaio de Thomas Nagel *What Is Like To Be A Bat*, em que uma crítica ao funcionalismo pode facilmente ser usada para criticar a Inteligência Artificial Forte. Outros filósofos, como Dreyfus, que veremos no segundo capítulo, miraram especificamente na Inteligência Artificial para fazer suas críticas, mas foi apenas John Searle que conseguiu fazer um estrago considerável.

John Searle talvez não seja um nome tão conhecido quanto Alan Turing, ao menos não fora dos círculos filosóficos, linguísticos e de ciência cognitiva. Nós iremos abordar o seu trabalho logo depois de Turing, porque sua crítica apesar de não ser a primeira, é provavelmente a mais

importante. Em seu artigo *Minds, Brains and Programs* Searle questiona a Inteligência Artificial Forte por sua ausência de intencionalidade, excessivo escoramento em analise comportamental, e, mais tarde, ausência de significado semântico. O que Searle fez em certa medida não foi algo necessariamente novo, o próprio Turing já havia se deparado com uma crítica em certa medida parecida quando escreveu seu ensaio. (A maneira como Turing lidou com objeções ao seu ensaio, veremos no subcapítulo a seguir). O ponto é que Searle foi muito feliz usando aquilo que tornou o ensaio de Turing famoso, que é um experimento. Ao propor seu teste, chamado *The Chinese Room,* Searle consegue condensar páginas inteiras de um texto filosófico em um teste que pode ser descrito, grosso modo, em quatro linhas, da mesma forma que Turing fez com o seu *Imitation Game.*[1]

A análise completa do trabalho de Searle está no segundo capítulo. A crítica de Searle foi tão forte que serviu para quebrar alguns paradigmas dentro da Inteligência Artificial. Primeiro, colocou em xeque o projeto de uma Inteligência Artificial Forte puramente simbólica, isto é, agentes inteligentes que se baseasse em regras lógicas de alto nível, buscas, redes e grafos para encontrar as suas soluções, colocando por terra toda uma geração de teóricos que defendia a IA Forte baseado em modelos simbólicos clássicos (chamados de GOFAI, ou *Gold-Old-Fashion-Artificial-Intelligence*), Segundo, serviu para que uma boa parte dos pesquisadores "deixassem a questão de lado", afinal, para a maioria dos pesquisadores tinha pouca importância saber se uma máquina de fato pensava ou não. A única forma de uma máquina de fato "pensar" era se ela não recorresse aos métodos clássicos da Inteligência Artificial. No caso, o agente precisava seguir um modelo de que simulasse a atividade neural biológica. Isso é visto dentro do conexionismo.

Falar de Conexionismo, o terceiro capítulo desse trabalho, implica falar de Redes Neurais, um assunto que facilmente foge do escopo filosófico. Não vamos entrar mais do que na superfície aqui. As Redes Neurais artificiais foram criadas na década de 1940 e tinham como objetivo uma simulação substancial da atividade neuronal biológica. No começo esse modelo de Inteligência

1 Daniel Dennett vai chamar esse tipo de experimento de *Intuition Pump,* que se trata de um experimento mental estruturado para permitir que o pensador use sua intuição para desenvolver uma resposta para o problema. Dennett descreveu o *Intuition Pump* tendo o *Chinese Room* em mente, apesar disso, no artigo *Computing Machinery and Intelligence* Turing chega a citar que uma pessoa que visse uma máquina passando no teste, ficaria inclinada a achar que ela de fato pensa, mesmo que isso contrafizesse as suas crenças pessoais. Mais do que isso, em seu artigo, Turing tentou descrever que mesmo que não houvesse nenhuma máquina que pudesse superar seu teste, ele já poderia ser considerado como uma possível evidência para futuras máquinas que pensam. Vamos assumir, nesse sentido, que o Teste de Turing também se trata de uma *Intuition Pump*. Mais detalhes em: Dennett, Daniel C. (2013). "Intuition Pumps and Other Tools for Thinking." New York: W. W. Norton & Company.

Artificial teve grande dificuldade de encontrar sucesso, pois as Redes Neurais Artificiais só eram capazes de reconhecer padrões simples, além de serem extremamente desafiadoras de se executar nos computadores da época. Apenas nos anos 1980 é que ressurge o interesse nesse tipo de tecnologia, bem como cresce também uma atenção da filosofia e da psicologia nessas redes neurais artificiais. O que estava em jogo era o fato de que uma rede neural artificial poderia simular o funcionamento da mente humana e nos dar respostas sobre questões como entendimento e a questão interna do funcionamento da mente. O trabalho dos filósofos Paul M. Churchland e Patricia S. Churchland é particularmente importante nesse sentido.

2. “COMPUTING MACHINERY AND INTELLIGENCE”

Quando se trata do problema de Máquinas com mentes, o Teste de Turing é o nosso primeiro ponto de partida. Nós analisaremos o trabalho original de Turing, bem como alguns poucos comentários feitos posteriormente pelos mais influentes acadêmicos da área, eles são muitos, mas veremos apenas os essenciais. A fama do artigo se deve porque Turing conseguiu, a sua época, ensaiar muito bem como uma suposta máquina com mente se comportaria, assim como também pode antever alguns questionamentos ao seu teste, também soube como propor avanços ao campo da engenharia da computação em geral, de forma que é possível analisar o artigo *Computing Machinery and Intelligence* sob múltiplos pontos de vista. Nesse sentido, muito material foi publicado, alguns se preocupando apenas com a questão cientifica, outros com foco teórico. O artigo original de Turing foi revisado e comentado em 2009 (Turing, 2009) por alguns dos seus estudiosos mais exaustivos, será essa a versão que usaremos aqui para debater o teste de Turing, bem como também usaremos o artigo de A. P. Saygin publicado no aniversário de 50 anos do Teste (Saygin, 2000). Vamos abordar e comentar apenas os trechos do teste que estão alinhados com a nossa problemática principal, no entanto, abriremos alguns parênteses paras certos tópicos de engenharia ligados a Inteligência Artificial, porque eles podem enriquecer a discussão, além de facilitar no entendimento das futuras descrições a respeito de como funcionam redes neurais artificias e programas de inteligência artificial em geral. Estamos falando de um campo que avança e que, querendo ou não, possuí maneiras diferentes de abordar os problemas se considerarmos o ano em que o Teste foi publicado, e o ano que estamos hoje.

Logo no começo do seu trabalho, Turing concebe a questão: "Podem as máquinas pensar?". Ele reconhece a abrangência da pergunta, de forma que propõe uma preocupação com o os termos “máquina” e “pensar”. Esse cuidado é uma evidencia de um olhar investigativo, apesar de que ele não busca o histórico dos termos, como algum filósofo faria (Há quem coloque Turing como filósofo, mas vamos nos livrar dessa problemática). Grosso modo, com “máquina” alguém pode se referir tanto a um computador com transistores, quanto a um motor de oito cilindros. “Pensar” não é menos ambíguo. Duas definições simples podem surgir a principio. Enquanto que alguém pode acreditar que “pensar” se trata de realizar equações matemáticas e lógicas, outro pode apontar que “pensar” diz respeito a um estado “consciente” e “senciente”, em que “existe alguma coisa acontecendo internamente”. O Jogo da Imitação, proposto por Turing serve justamente para esclarecer esses conceitos, e assim como nos capítulos posteriores do artigo, Turing deixa claro que com “máquina”, ele não quer dizer uma máquina banal, como um motor a combustão de oito

cilindros, mas sim uma máquina com a capacidade de computar. Essa diferenciação é importante de se entender porque entre "máquinas pode pensar?" e "computadores podem pensar?" alguém pode dizer que existe uma diferença significativa. Uma máquina é qualquer coisa em que haja movimento mecânico artificial. Uma turbina eólica é uma máquina, mas uma turbina eólica não é um computador. Usa-se a definição "máquina" por uma questão de comodismo, mas Turing vai dizer que com "máquina", ele quer dizer "máquina de estados discretos", ou seja, um computador. Portanto, com "máquinas podem pensar?", estamos perguntando, "computadores podem pensar?".

Voltando a nossa discussão sobre a definição de pensar levantada por Turing. Este caso está mais a favor da definição de que pensar diz respeito a existência de estados mentais internos. Por padrão assumimos pensar como dois tipos de atividade, uma é aquela ligada ao raciocínio, isto é, resolver problemas, e a outra está ligada ao domínio dos estados mentais, que no caso quer dizer que não importa se estamos resolvendo cálculos de matemática ou não, se estivemos ocupados com alguma coisa, ou mesmo ociosos, estamos "pensando" (É a definição o que se verifica em consenso entre filósofos e cientistas). "pensar" nesse sentido é um estado mental interno, em certo sentido relacionado a consciência. Pode ser descrito também como "a consciência da execução de tarefas", e é com esse "pensar" que Turing está preocupado, de acordo com aquilo que é descrito no seu artigo. Mesmo que o uso do termo "consciente" possa implicar um caminho espinhoso, pelo fato de que Turing não assume que pensar e consciência são a mesma coisa[1]. Que queremos dizer é que com "pensar", estamos falando de algo que "é mais do que um objeto inanimando e pode pensar a respeito das tarefas que executa". Nesse sentido, a partir dessas evidências, podemos redefinir a frase "Máquinas podem pensar?", como "computadores tem estados mentais internos?", sem medo de cometer algum equívoco em relação a aquilo que Turing originalmente pretendia.

Quanto ao teste, Turing é claro a respeito daquilo que quer, no caso um teste comportamental, apesar de que em momento algum ele o descreve dessa forma. Turing, porém, vai afirmar mais de uma vez que o projeto do *Jogo da Imitação* é a avaliação do comportamento e que através dele nós poderíamos dizer se uma máquina pensa ou não, não importando, inclusive, como a máquina foi programada. O que importa para Turing é saber se a máquina é capaz de pensar no teste e não de observar estados internos ou uma possível atividade neural. Isso nos coloca numa posição do debate que pode vir a ser expandida futuramente, pois afinal, que relevância tem a construção interna de uma máquina caso ele passe no teste? para Turing, nenhuma. Porém, os filósofos da mente vão

1 Turing aborda a questão da consciência em resposta a uma das objeções ao seu teste. Ele descreve que "não precisamos resolver os mistérios da consciência para dizer se uma máquina pensa ou não." O que leva a crer que Turing considerava "consciência" e "pensar" como duas coisas distintas.

colocar essa questão em discussão, especialmente porque, baseado naquilo que Turing descreveu, para construir uma máquina que passe no teste, era preciso se basear em símbolos. Segue o resumo de como funciona o *Jogo da Imitação* (mais tarde chamado de Teste de Turing)[1].

O jogo é jogado com um homem (A), uma mulher (B) e um interrogador (C), cujo o sexo não é importante. O interrogador fica em uma sala separada de A e B. O objetivo do interrogador é determinar qual dos outros dois é a mulher enquanto o objetivo de ambos, homem e mulher, é convencer o interrogador de que ele é a mulher e o outro não. Essa situação está representada na Figura 1. O meio pelo qual a conversa deve ocorrer é uma conexão de teletipo. Assim, o interrogador faz perguntas em linguagem natural escrita e recebe respostas em linguagem natural escrita. As perguntas podem ser sobre qualquer assunto, desde matemática a as artes. Nesse sentido, caso passe no teste, a pergunta que se faz é "A(s) máquina(s) pode(m) pensar?"

Turing usa a definição do seu teste como "O Jogo da Imitação", mas o fato é que há pelo menos mais de um teste, desdobrado a partir de uma configuração inicial. Inclusive, havendo a possibilidade de adaptar o teste para situações especificas.

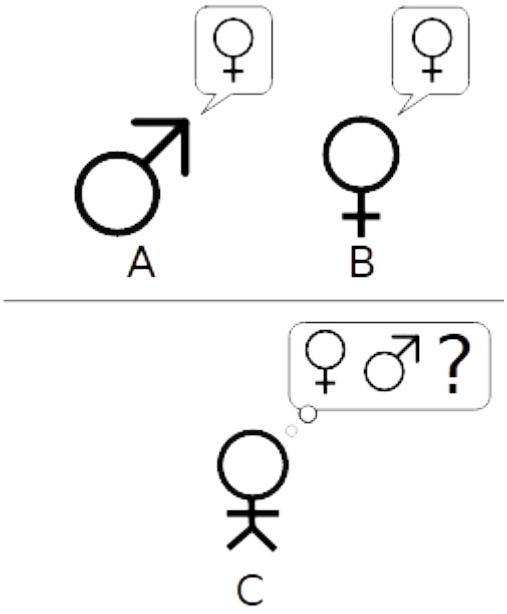

Figura 1: Exemplo do Teste de Turing

1 Turing não chama sua ideia de "Teste de Turing", mas sim de "Jogo de imitação"; no entanto, a literatura posterior reservou o termo "Jogo de imitação" para descrever uma versão específica do teste, que no caso é versão em que a máquina tem que se passar por uma mulher/humano. (Há, pelo menos, três versões, já que Turing modifica o teste em várias situações.);

Uma pessoa conversando (dentro do experimento) com um robô por meio de Mensagens, por exemplo, pode ser considerado um Teste de Turing, apesar de que os mais puristas tendem a usar apenas o primeiro modelo proposto, onde a máquina precisa se passar por uma mulher e há um terceiro interrogador. Claro, é preciso conceber que o teste de Turing tem regras, mas o próprio Turing flexibiliza essas regras ao longo do seu ensaio e também dependendo do contexto.

No que diz respeito ao nome do teste "Jogo da Imitação", o comentador Stevan Harnad (Harnad, 1992) pontua que chamá-lo dessa forma foi um ato falho por parte de Turing, uma vez que se trata de um mais teste mais sério do que parece. Diz ele: "Chamar o experimento de "jogo de imitação" em vez de uma metodologia para a capacidade de desempenho cognitivo humano de engenharia reversa resultou em gerações de mal-entendidos desnecessários" (Harnad, 1992). Vamos levar em conta que seja um tanto quanto forçoso assumir que Turing daria um nome como esse, uma vez que não existia no tempo em que ele escreveu uma ciência cognitiva propriamente dita, além do método behaviorista. É verdade que a analise comportamental dessa época é muito mais limitada do que as ciências cognitivas que temos hoje, inclusive, o Jogo da Imitação é uma dessas análises relativamente simples, mas que se propõe a resolver uma questão muito profunda. Podemos imaginar que Turing chama o seu teste de "Jogo da Imitação" mais por estar propondo algo novo e até então difícil de se classificar na psicologia do seu tempo. Em seu ensaio, Turing não esclarece o porquê da nomenclatura "Jogo", a ideia inicial é que ele o nomeou dessa forma mais por uma trivialidade, é um jogo com regras, onde uma máquina deve "imitar" um ser humano e caso consiga, acaba vencendo.

Como um engenheiro, Turing sabia da possível falta de necessidade prática para o seu teste. Computadores até o seu tempo eram máquinas úteis, tinham um propósito de servir a guerra e, mais tarde, a ciência e aos negócios. Ainda assim, posterior a Turing, a pesquisa na ciência da computação sofreu muitos reveses no que diz respeito a credibilidade em relação a Inteligência Artificial. Nesse sentido, Turing pergunta em seu artigo "*Is this new question a worthy one to investigate?"*. Já consciente de que algum crítico poderia questionar sobre a ausência de utilidade para o caso. Ele não fornece uma resposta direta, mas argumenta que investigar a possibilidade de máquinas possuírem pensamento não é a mesma coisa que dizer se elas são como nós, um suposto antropomorfismo. A ausência de um argumento mais claro de Turing para essa questão evidência o seu caráter mais ontológico. Mesmo hoje em dia, com o avanço das ciências cognitivas e da Inteligência Artificial, tentamos encontrar usos práticos para programas de processamento de linguagem natural, robôs humanoides e tantas outras máquinas que nós fazemos com o propósito de

copiar a nós mesmos. Foi pensando nisso que Turing demonstrou como supostamente funcionaria a comunicação entre máquina e ser humano. Ele elenca alguns pontos interessante da conversação e exemplifica no ensaio dessa forma:

1. Pergunta: Por favor, escreva-me um soneto sobre o tema da Ponte Forth.

   Resposta: Me deixe fora dessa. Nunca consegui escrever poesia.

2. Pergunta: Adicione 34957 a 70764.

   Resposta: (Pausa cerca de 30 s e depois dá como resposta) 105621.

3. Pergunta: Você joga xadrez?

   Resposta: Sim.

4. Pergunta: Eu tenho K no meu K1 e nenhuma outra peça. Você tem apenas K em K6 e R em R1. É o seu movimento. O que você toca?

   Resposta: (Após uma pausa de 15 s) R-R8 mate

Vamos fazer um pequeno exercício comentado esse diálogo, mesclando ciência e filosofia. É possível construir um software que seja capaz de ter uma conversa como essa usando apenas IA Simbólica (originalmente, a Inteligência Artificial era inteiramente simbólica). Em P1-R1 temos provavelmente o ponto mais discutível dessa conversa. É verdade, do ponto de vista da engenharia, que nenhum programa de IA Simbólica pode produzir alguma coisa que seja considerada "artística", como, por exemplo, escrever poesia. A perspicácia de Turing em considerar essa questão é admirável, mas mais admirável ainda é perceber que com isso ele praticamente assume que uma máquina não precisa produzir nada para pensar. O que podemos tirar disso? Grosso modo, nenhum animal além do ser humano possuí capacidade cognitiva para se expressar da maneira como nós fazemos. Nesse sentido Turing assume que uma máquina não precisa chegar ao nosso "nível mental" para que seja possível assumir que ela pensa. Poderíamos desdobrar desse ponto da conversa a questão "máquinas produzem arte?", ao qual Turing, com esse pequeno exemplo, assume que não. Do ponto de vista da engenharia, mesmo que hoje em dia um programa usando Redes Neurais Artificiais consiga construir imagens originais a partir de exemplos e treinamentos, uma rede neural treinada conseguiria de alguma maneira produzir (copiar) uma poesia, mas tanto os artistas, quanto os filósofos, tem bons motivos para não chamar o que sai de dentro de uma máquina de "arte". Abordaremos isso com mais detalhes no segundo capítulo.

Sobre P2-R2. Qualquer programador ou engenheiro competente pode escrever um programa que resolve cálculos a partir do processamento de linguagem natural (“Qual a raiz quadrada de 2?”, “Quanto rende cem mil reais aplicados em dois anos de Tesouro Selic, considerando que o Tesouro rende 11% ao ano?”). Em Q3-A3 se trata de unir duas funções diferentes no mesmo programa. Isso é totalmente possível com IA Simbólica. Ademais, se trata de uma das ideias da Inteligência Artificial Geral, unir várias funções num programa, que acaba sendo um “faz-tudo”. Em Q4-A4: Para interpretar isso, o programa usaria o mesmo módulo de funções que usa para interpretar “Adicione 34957 a 70764”. Nesse sentido, temos o *Deep Blue* como exemplo de máquina que derrotou um ser humano dessa mesma maneira, usando inclusive IA Simbólica. Por fim, antes de irmos as críticas respondidas por Turing ao seu teste, é preciso destacar que Turing considera que o método acima descrito, isto é, o método de comunicação “Pergunta-Resposta”, *(Question-Answering Program*) é a melhor maneira de implementarmos uma comunicação natural entre homem e máquina. Nesse sentido, podemos observar que praticamente todos os programas relevantes que pretenderam passar no teste e outros programas mais modernos de ciência cognitiva são *Question-Answering Programs.* Alguns exemplos podemos elencar como ELIZA (Weizebaum, 1966), PARRY (Colby, 1972), o supercomputador WATSON, da IBM e mais recentemente, o ChatGPT lançado pela OpenAI em 2022.

Uma questão aos quais os comentadores de Turing tratam, bem como uma das objeções que o ensaio recebeu (P.H. Millar, 1973) diz respeito a natureza das perguntas e das respostas. Alguns comentadores acusam Turing de antropomorfismo (apesar de que, como vimos, ele declina essa crítica), e outros questionam a validade das perguntas, tratando-as como sendo insuficientes para se avaliar o pensamento ou inteligencia. Uma coisa é alguém perguntar: “Me diga qual a raiz quadrada de cento e vinte e oito, e depois dívida por dois”, mais ou menos como na pergunta P2 do dialogo descrito por Turing. Outra coisa é alguém perguntar: “Tenho a sensação de que estou vivendo uma vida muito monótona, não saio de casa há duas semanas, o que você acha?”. São perguntas de ordem diferente e que exigem níveis de sofisticação diferentes para a resposta. Enquanto a primeira está mais para uma questão puramente matemática, a segunda exige um certo conhecimento de mundo da máquina, algo que Turing considera que a máquina deve possuir, mas que é importante de ser ressaltado porque quando se trata de um diálogo baseado em perguntas e respostas, não estamos falando meramente de um programa que responde questões matemáticas e lógicas, estamos falando de um programa que é capaz de dar respostas a problemas mais abstratos baseado no seu conhecimento e aprendizado interno.

2.1 VISÕES CONTRÁRIAS

Em seu artigo original, Turing listou nove argumentos que poderiam ser usados contra o seu Jogo. Desses nove, vamos destacar apenas três aqui, pois são esses os que estão mais de acordo com aquilo que queremos pesquisar. Esses três argumentos, Turing chamou respectivamente de "Objeção Teológica", "Objeção Matemática" e "Argumento baseado em Consciência".

> Acredito que em cerca de cinquenta anos será possível aos computadores, com uma capacidade de armazenamento adequada, jogar o jogo da imitação tão bem que um interrogador médio não terá mais de setenta por cento de chance de fazer a identificação correta após cinco minutos de interrogatório. - (TURING, 1950, p. 442)

2.1.1 OBJEÇÃO TEOLÓGICA

A objeção teológica (que também pode ser chamada de objeção da imaterialidade da alma) diz que o pensamento é uma das funções da alma imortal criada por Deus, exclusiva do homem e que nenhum animal ou máquina pode possuir aquilo que é concedido apenas por Deus ao homem. Esse argumento é minimamente semelhante ao dualismo cartesiano, especialmente na parte que assume que os animais não possuem alma (segundo essa visão, eles seriam autômatos). Também podemos acreditar que se um dualista se deparasse com o Teste de Turing, provavelmente seria um argumento parecido com esse que viria a usar. Em todo caso, se trata de um argumento bastante generalista, que poderia ter um derivado semelhante sendo usado tanto por um religioso quanto por um intelectual não-reducionista, por exemplo. Turing não responde esse argumento a moda materialista, apesar de que ele poderia. É interessante contextualizar que ao mesmo tempo em que Turing publicava o seu artigo, saia, um ano antes, o trabalho de Gilbert Ryle, que entre outras coisa, recusaria o problema "mente-corpo" e daria início a filosofia da mente como uma abordagem analítica, bem como ao behaviorismo filosófico.

A réplica de Turing é mais ou menos dividida em dois argumentos. Primeiro ele se recusa a acreditar que os animais, com toda a sua complexidade comportamental, sejam meros autômatos sem alma, e busca uma saída teológica para a sua argumentação. Diz ele que, afinal, se Deus é o todo-poderoso, ele teria o poder de conceder uma alma a algum animal que tivesse um cérebro bem desenvolvido o bastante para "receber" uma alma, assim como uma máquina, caso fosse complexa o suficiente, poderia receber algo semelhante. Mais do que isso, Turing questiona o ponto de vista mais doutrinário dessa argumentação, já que dentro da cultura islâmica, existe a crença de que apenas pessoas do sexo masculino possuem alma, o que para um cristão pode ser absurdo. Por

último, usa um argumento mais provocativo, relembrando o caso em que algumas passagens da bíblia serviram para sustentar visões de mundo hoje consideradas ultrapassadas, como o caso do geocentrismo. Nesse sentido, assim como o descobrimento do heliocentrismo, a mente nas máquinas seriam uma questão de avanço científico, não tendo nada a ver com um atributo concedido por Deus.

2.1.2 OBJEÇÃO MATEMÁTICA

A objeção matemática se vale do argumento de Incompletude de Gödel[1], que podemos resumir como o caso de que em qualquer sistema lógico suficientemente poderoso podem ser formuladas declarações que não podem ser provadas nem refutadas dentro do sistema, a menos que possivelmente o próprio sistema seja inconsistente. Dentro dessa lógica, pode ser traduzida, de uma maneira mais informal, a objeção: "Existem problemas que as máquinas nunca poderão resolver, portanto elas não podem ser inteligentes". Essa objeção ataca não apenas a possibilidade de máquinas com mentes, mas a Inteligência Artificial como um todo, pois aqui não estamos lidando apenas com o caso de que máquinas podem pensar ou não, mas o caso de que se elas são inteligentes ou não. Mais para frente, no capítulo seguinte, veremos que Dreyfus faz uma crítica semelhante a Inteligência Artificial.

Turing resolve essa objeção com uma resposta bastante convincente, forte o bastante para resistir até hoje (talvez por isso poucos vieram a relembrar da objeção matemática). O fato de existir problemas que máquinas não podem resolver não significa que elas sejam incapazes no geral. Assim como seres humanos dificilmente podem resolver cálculos complexos usando apenas a mente, ou mesmo serem intelectualmente incapazes de mensurar certas grandezas, não significa que haja ausência de inteligência. A mente humana, assim como as máquinas, possui suas limitações. Mais do que isso, em seu teorema, Godel não prova que seja impossível encontrar a verdade, apenas aponta que a verdade não pode ser provada. Nesse sentido também é possível imaginar que "pensar" não é o mesmo que "provar".

---

1 Os teoremas da incompletude de Gödel são dois teoremas da lógica matemática que se preocupam com os limites da provabilidade em teorias axiomáticas formais. Esses resultados, publicados por Kurt Gödel em 1931, são importantes tanto na lógica matemática quanto na filosofia da matemática. Os teoremas são amplamente interpretados como evidências que o programa de Hilbert para encontrar um conjunto completo e consistente de axiomas para toda a matemática é impossível. Ver: Kurt Gödel – Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/entries/goedel/) *Acesso em 2022*

2.1.3 ARGUMENTO BASEADO EM CONSCIÊNCIA

Esse é provavelmente o argumento mais importante e mais duro que o Turing considerou. Ele também é muito parecido com as críticas que a Inteligência Artificial Forte vai receber a partir dos anos 1970. Seu autor é Geoffrey Jefferson, que foi um neurocirurgião pioneiro em diversas áreas da sua ciência, além de grande contribuinte para a pesquisa de cirurgia médica em geral. Jefferson foi honrado em 1948 com a *Lister Medal*, distinção dada pelo *Royal College of Surgeons of England* a médicos que contribuíram de alguma forma a ciência cirúrgica. Ao receber a honraria, Jefferson ministrou uma *Lister Oration*, uma aula em agradecimento ao prêmio. Jefferson publicou sua *Lister Oration* em 1949, sob o título *The Mind of Mechanical Man,* onde ele promoveu um debate sobre a possibilidade de Inteligência Artificial. Turing teve contato com essa publicação, em seu artigo ele mesmo destacou o ponto mais importante do texto de Jefferson, que se segue:

> Quando uma máquina puder escrever um soneto ou compor um concerto por causa de pensamentos e emoções sentidas, e não pelo uso casual de símbolos, poderemos concordar que a máquina é igual ao cérebro - isto é, não apenas escrever algo, mas saber que o escreveu. Nenhum mecanismo pode sentir (não apenas sinalizar artificialmente, que é um artifício fácil) prazer com seus sucessos, ficar aflito quando suas válvulas se fundem, ser aquecido por comentários lisonjeiros, ficar incomodado por seus erros, ser encantado pelo sexo, ficar zangado ou deprimido quando não pode conseguir o que quer. (JEFFERSON, 1949, p. 211)

Dentro desse trecho nós podemos derivar uma variedade grande de argumentos, dentre os quais o principal é o Problema das outras mentes (*Other Minds*). Esse também é a única das nove objeções as quais Turing não dá uma resposta realmente satisfatória. Primeiro ele assume que esse argumento se trata de uma recusa ao Teste de Turing, o que é verdade, se seguirmos a partir da segunda linha de pensamento desse trecho. Mas além disso, é também um recusa ao qualquer teste que tenha um caráter behaviorista como o de Turing.

Qual a resposta de Turing para isso? “Para o Teste funcionar é preciso aceitar o Teste”. Parece um argumento simplório, mas se levarmos em consideração que de fato Jefferson não leva o teste a sério, sob a linha de raciocínio de que para pensar uma máquina precisa fazer tudo o que nós fazemos (Destaque para “Prazer com seus sucessos, tristeza quando suas válvulas se fundem, ser aquecido por lisonjas, ser miserável por seus erros, ser encantado pelo sexo, ficar com raiva ou deprimido quando não consegue o que quer”) incluindo eventos emocionais, faz todo o sentido imaginar que ele não tome o teste como válido. O argumento que de para ser inteligente (ou pensar) uma máquina precisa fazer tudo aquilo que fazemos também foi explorado posteriormente por Keith Gunderson. Em resumo, Gunderson assumiu que o teste não poderia ser válido porque “imitar apenas uma característica humana como uma conversa não é o suficiente para dizer se algo ou

alguém pensa ou não" (Gunderson, 1964). Além do caso da não aceitação do teste, há também o caso do Problemas das Outras Mentes, ao qual Turing acredita que partir do pressuposto que para saber se *x* pensa é preciso ser *x* se trata de uma visão solipsista, já que segundo raciocínio de Turing, alguém que acredita nisso anula a existência de todos os outros seres pensantes, assumindo a apenas a própria existência. Stevan Harnad (Harnad, 1991), aponta que ao assumir o problema *Other Minds* como "solipsismo", Turing comete um erro crasso. Uma resposta convincente, segundo Harnad, seria de que, apesar de existir verdade no problema *Other Minds*, estamos falando de uma análise de comportamento, e nesse sentido, tomando a nós como exemplo, ninguém dúvida que qualquer pessoa parecida conosco possa pensar, pelo fato de sua consistência e seu comportamento ser semelhante ao nosso.

Há outros pontos interessantes sobre esse trecho do texto de Jefferson que podemos discutir, especialmente as primeiras frases "Quando uma máquina puder escrever um soneto ou compor um concerto por causa de pensamentos e emoções sentidas, e não pelo uso casual de símbolos*"* e também "poderemos concordar que a máquina é igual ao cérebro - isto é, não apenas escrever algo, mas saber que o escreveu*"*. Nesse primeiro trecho temos um argumento que John Searle irá repetir. Trata-se do caso de que a manipulação de símbolos não é suficiente para conceder pensamento a uma máquina. Esse é um ponto importante pois toca na questão da construção interna das máquinas (algo que Turing ignora. Já que para ele não interessa como uma máquina funciona internamente, a única coisa que importa é se ela passa no teste!). O segundo trecho diz respeito a consciência. Turing entende que a consciência é importante e não deve ser tratada de maneira leviana, mas ele acredita, como já vimos, que o fato de não termos respostas para todas as questões da consciência, bem como sabermos ou não se internamente uma máquina é capaz de entender o que está fazendo, não deve ser um obstáculo para a pesquisa de máquinas com mentes. Essa linha de pensamento é bastante semelhante ao funcionalismo.

Depois de apontar o problema das outras mentes como um problema solipsista, Turing faz um movimento interessante. Digamos que ele torna o teste mais "moderno" ao eliminar o jogador B do Jogo da Imitação e considerar o jogo agora não mais como uma imitação, mas como um *viva voce*[1]

1 *Viva Voce* é basicamente a transmissão do conhecimento oralmente. Aos filósofos a referência é a maiêutica, apesar de que o *viva voce* está mais relacionado a tradição oral do *folk-lore*. No caso do Teste de Turing faz sentido compreendê-lo como uma "maiêutica", pois o que se faz é o exercício de questionar uma máquina para saber se ela realmente compreende não apenas a comunicação do ser humano, com todas as suas nuances, mas também um assunto específico, em vez de simplesmente apanhar conhecimentos avulsos de um banco de dados. É um fato peculiar que dificilmente as *virtual assistants* de hoje passariam no *viva voce* de Turing, pelo fato de que elas basicamente fazem aquilo que Jefferson descreveu: Apanham uma informação de um banco de dados e apresentam ao usuário.

(uma conversa sobre um determinado assunto). Chamamos essa forma do jogo de mais moderna porque ela é mais fácil de se replicar hoje em dia, basta se ter um programa e um usuário disposto a testar a máquina. Podemos transcrever o conteúdo da conversa que Turing exemplificou:

> Interrogador: Na primeira linha do seu soneto que diz "Devo te comparar a um dia de verão", "um dia de primavera" não seria tão bom ou melhor?
> Testemunha: Não tem a métrica certa
> Interrogador: Que tal "um dia de inverno" Isso tem a métrica.
> Testemunha: Sim, mas ninguém quer ser comparado a um dia de inverno.
> Interrogador: Você diria que o Sr. Pickwick o lembrou do Natal?
> Testemunha: De certa forma.
> Interrogador: No entanto, o Natal é um dia de inverno, e não acho que o Sr. Pickwick se importaria com a comparação.
> Testemunha: Não acho que você esteja falando sério. Por esfola de inverno, entende-se um dia típico de inverno, em vez de um dia especial como o Natal. (TURING, 1950, p. 446)

O prêmio Loebner é um prêmio dedicado a programas que possam passar no Teste de Turing, apesar de alguns programas terem alcançado um nível de conversação semelhante a esse, nenhuma pode chegar a esse nível de profundidade em uma conversa. Acadêmicos também criticam o prêmio por ser muito superficial. O ponto é que, se uma máquina conseguisse chegar a esse nível de conversação, Turing acredita que Jefferson ficaria extremamente convencido que a máquina é realmente capaz de pensar, independentemente da construção interna do programa. Dissemos que Turing não fez uma réplica satisfatória a essa objeção específica, e isso ocorre justamente por que aqui todas as tentativas argumentativas de Turing baseiam-se em "você deve aceitar o teste para ele funcionar" ou "se visse uma máquina conversando com um humano dessa forma, você certamente se convenceria". O problema é que, pelo menos para a filosofia, uma máquina se comportar como um ser humano, conversar como um ser humano ou parecer pensar como um ser humano não basta. As aparências não são suficientes para a investigação filosófica e ainda que Turing não seja um filósofo no sentido mais rigoroso do termo, ele deveria ter considerado esse caso.

É nesse sentido que as críticas mais pujantes ao Teste de Turing e as máquinas com mentes em geral vão se delinear e são essas críticas que veremos no capítulo seguinte. Antes vale lembrar que Turing em seu artigo considera ao menos nove objeções, aos quais não tratamos aqui pelo fato de que estamos mais focados com o problema das máquinas com mentes do que com o problema do Teste de Turing. Esse não é um trabalho exclusivamente sobre o teste, apesar da importância dele.

Também é importante ressaltar que existe toda uma literatura em volta do teste de Turing, os interessados devem buscar A. P. Saygin (Saygin, 2000) bem como Harnard (Harnard, 2004).

3. DREYFUS, SEARLE E A CRÍTICA A MÁQUINAS COM MENTES.

Entre 1950 e 1964 a ideia de máquinas inteligentes desenvolveu-se sem maiores divergências filosóficas. No capítulo anterior destacamos a diferença que existe entre "Inteligência Artificial Forte" e "Inteligência Artificial Fraca". Essa diferença serve para explicar quando alguém, por exemplo, fala de um agente inteligente que pode vencer um jogo de Xadrez e sobre quando alguém fala de um agente inteligente que pode *pensar.* Quando o filósofo Humbert Dreyfus publicou em 1972 seu trabalho "*What Computer's Can't Do"* essas distinções não existiam, e inteligência artificial poderia significar tanto um programa que pode resolver um quebra-cabeças de 9 peças, quanto um programa capaz de manter uma conversa que pudesse imitar um ser humano real. Por causa disso, Dreyfus, o primeiro crítico severo da Inteligência Artificial, não tinha limites quanto a sua crítica à pesquisa em IA. Para ele, qualquer forma de inteligência artificial era impossível. Seria tão inconcebível uma máquina pensar quanto demostrar o mínimo de capacidade lógica para resolver problemas como um ser humano ou qualquer coisa que pudéssemos chamar de inteligência. Discutimos no capítulo anterior que, sob certos aspectos, é razoável supor que uma máquina seja inteligente, porém quando se trata de *pensamento* a questão se torna um tanto mais complicada. Pois bem, quando Dreyfus fala sobre a impossibilidade de inteligência artificial ele fala tanto da incapacidade de uma máquina se usar raciocínio lógico quanto a de se passar por um ser humano (Dreyfus, 1972).

Humbert Dreyfus foi um filósofo norte-americano especializado em filosofia continental, em especial, fenomenologia e existencialismo. Devotou grandes esforços na análise do trabalho de Martin Heidegger e Merleau-Ponty, seu trabalho mais relevante nessa área é sua tese *Husserl's Phenomenology of Perception* de 1964. Apenas saber que Dreyfus dedicava interesse em especial a fenomenologia já deve dar algumas pistas do porque ele teria algum eventual problema com a Inteligência Artificial. Sua primeira crítica a ideia de agentes inteligentes surgiu em 1965, em formato de um memorando que produziu como pesquisador consultor da RAND, mas o trabalho de real relevância só veio a ser publicado em 1972. Nesse trabalho, em resumo, Dreyfus explica que a "filosofia" da Inteligência Artificial está escorada em fundamento filosófico que podemos chamar de behaviorista. Para ele, os cientistas e engenheiros envolvidos no projeto da IA acreditam que podem, assim como nas ciências naturais, condensar e descrever o comportamento humano por meio de regras, tal qual as regras de um movimento planetário. Ao logo da sua crítica, Dreyfus escreveu quatro livros sobre o tema, o original que já citamos de 1972, uma reedição de 1979, um terceiro, este chamado de *What Computers Still Can't Do* de 1992, e *Mind Over Machine,* de 1986.

Apenas o que está no primeiro livro e o resumo do argumento dos outros três nos interessa nesse momento. Antes de explicar a critica de Dreyfus de maneira concisa, é importante destacar o que a Inteligência Artificial viveu nesse período, de 1950 a 1972.

Originalmente, a data escolhida para a fundação da Inteligência Artificial é 1956, após a famosa conferência no Dartmouth College. Como já destacamos antes, houve uma cadeia de acontecimentos para que isso acontecesse, entre os quais, a publicação de *Computing Machinery and Intelligence*. Quando os primeiros programas de inteligência artificial começaram a apresentar algum sucesso, a euforia entre os cientistas foi tão grande que chegava a ocupar algum espaço na mídia tradicional. Os cientistas John McCarthy, Allen Newell e Hebert Simon são alguns dos nomes que fizeram parte desse processo. Entre outras coisas, eles conseguiram criar programas que foram muito bem sucedidos, porém de forma limitada. Tomemos como exemplo um programa de Newell e Hebert Simon (1959), o General Problem Solver (GPS). O objetivo do GPS era funcionar como um resolvedor geral de problemas (no caso, problemas lógicos). A ideia do GPS era trabalhar com um programa que armazenasse algum conhecimento e fosse capaz de acessá-lo posteriormente para resolver problemas. Tecnicamente, o GPS é descrito como um agente inteligente baseado em conhecimento, porque todos os problemas que ele precisa resolver já estão em sua biblioteca de dados. Esse é um dos casos de um programa mais técnico da época que foi bem sucedido, mas temos casos como o do pesquisador Arthur Samuel, que escreveu em 1956 um programa capaz de jogar damas de maneira satisfatória e que foi demonstrado na televisão. Em certo sentido, os sucessos eram grandes para a época, e cabe lembrar que estamos falando de um tempo em que o transistor acabava de ser inventado e um computador não era mais do que uma maquina de calcular, grande e extremamente difícil de ser programado. Essa impressão de programas resolvendo por si mesmo quebra-cabeças ou sendo capazes de jogar damas causou fortes expectativas, ao mesmo tempo, os pesquisadores não poupavam palavras ao fazer previsões ousadas. Se Alan Turing estimou um tempo de 50 anos para que os computadores fossem capazes de passar no seu teste, os pesquisadores acreditavam que em 10 anos um computador já venceria um g*randmaster* do xadrez (Isso só viria a ocorrer mais tarde, 40 anos depois). Hebert Simon, provavelmente um dos mais respeitados cientista da computação da história, além de notável cientista político, deu declarações bastante entusiasmadas sobre os avanços na época.

> Não é meu objetivo surpreendê-los ou chocá-los, mas o modo mais simples de resumir tudo isso é dizer que agora existem no mundo máquinas que pensam, aprendem e criam. Além disso, sua capacidade de realizar essas atividades está crescendo rapidamente até o ponto — em um futuro visível — no qual a variedade de problemas com que elas poderão lidar será correspondente à variedade de problemas com os quais lida a mente humana. (RUSSELL & NORVIG, 2004, p. 45).

Simon, por bem ou por mal, acabou sendo mais reconhecido por essa frase do que pelo seu trabalho como cientista da computação. Em entrevista anos mais tarde ele falou um pouco sobre a euforia dos pesquisadores da IA e a sua própria euforia.

> Se você olhar para outras ciências, que talvez não sejam tão pessoalmente ameaçadoras para as pessoas, previsões são feitas o tempo todo. Veja os cânones de comportamento na astronomia hoje. Você sabe, alguém pode andar por aí com a menor centelha de evidência e argumentar sobre novo tipo de universo que se expande ou contrai ou está permanentemente em um estado ou outro. Os cosmólogos fazem isso o tempo todo e são considerados bons cientistas em astronomia porque isso faz parte dos costumes desse campo. (McCORDDUCK, 2004, p. 222).

Fato é que esse entusiamo e essas afirmações ousadas viriam a assombrar a inteligência artificial por muito tempo. Quando publicou seu artigo original em 1965, Dreyfus tinha exatamente essas previsões em mente. Era, para ele, audacioso demais que os pesquisadores de IA fossem capazes de dizer coisas tão levianas sobre algo tão importante quanto a mente e o pensamento humano. O caso curioso para Dreyfus é que dentro da academia norte-americana, sua linha de pesquisa filosófica (ao menos na época) era um tanto quanto "atípica". Enquanto a maioria dos filósofos envolvidos com as questões da computação eram analíticos e estavam mais para o lado da linguagem, Dreyfus vinha com um conhecimento que naquela época era tratado como "alternativo" pela academia cientifica norte-americana (McCordduck, 2004). Por exemplo, a maioria dos cientistas da IA que tinham treinamento em filosofia eram especializados em filosofia da linguagem ou filosofia da ciência. Até a chegada do trabalho de Dreyfus não há praticamente nenhum comentário fenomenologista sobre o Teste de Turing, e apesar de contemporâneos ao desenvolvimento da Inteligência Artificial, nomes como Sartre, Merleau-Ponty e Heidegger não produziram nenhum grande comentário sobre o tema.[1] Nesse sentido podemos dizer que Dreyfus foi a primeira grande crítica da filosofia continental no que diz respeito à Inteligência Artificial que se fazia na época. Em um exercício imaginativo (e levando o trabalho de Dreyfus em consideração) era como se Merleau-Ponty resolvessem dirigir as suas atenções para o que a ciência da computação fazia naquele momento. Na Introdução do seu

1 No prefácio de *Phénoménologie de la perception* (1945) Merleau-Ponty assume uma posição critica a ciência cognitiva, ao qual ele descreve como "ingênua" a tentativa da compreender a mente por meios empíricos. O trabalho de Merleau-Ponty influenciaria Dreyfus em sua crítica contra a Inteligência Artificial e posteriormente o chamado "Pós-cognitivismo", do qual Dreyfus é um dos representantes.

livro *Mind Over Machine* (1986) Dreyfus comenta algo a respeito. No inicio dos anos 1960, segundo ele, era comum durante suas tradicionais aulas de filosofia no MIT os alunos comentarem sobre os avanços que a computação fazia e como em pouco tempo aquela filosofia seria algo do "passado". Dreyfus sabia que tinha pouco conhecimento e precisou adentrar no mundo da pesquisa em IA como consultor na RAND. Sua crítica pode ser dividida em dois pontos, aos quais comentaremos detidamente. Essa divisão não está explícita na sua obra, no entanto, pelo o que percebemos, existem um ponto do comentário de Dreyfus que concerne especialmente a máquinas com mentes. Essas críticas são: 1°). Existem aspectos da inteligência humana que uma máquina nunca seria capaz de copiar por meio de regras; 2°). Seres humanos são dotados de intencionalidade, ao contrário do que os naturalistas querem, a mente humana não pode ser analisada através do aspecto analítico;

Em relação ao primeiro ponto, Dreyfus tratou com veemência no seu primeiro trabalho a respeito do jogo de xadrez. A previsão, em 1956, era que em 10 anos um computador fosse capaz de vencer um g*randmaster*. Isso não só não aconteceu, como também eram poucos os programas que eram capazes de jogar xadrez a ponto de vencer uma pessoa minimamente treinada no Jogo. A primeira coisa que Dreyfus fala a respeito, em seu trabalho de 1972, é sobre a incapacidade que o programa teve em vencer um jogador de 10 anos de idade em uma partida organizada pelos pesquisadores como parte da divulgação do projeto da IA. Para Dreyfus, entretanto, existia um motivo para isso. Pelo fato do comportamento humano ser complexo demais, a abordagem por *regras,* a qual Dreyfus assumiu que os pesquisadores de IA seguiam, nunca seria capaz de gerar um comportamento artificial tão imprevisível, tão complexo e tão irredutível como o comportamento humano. Essa inabilidade em reunir todo o conjunto de regras lógicas é chamada de problema da qualificação. Tal argumento já havia sido esboçado por Turing em 1950, mas Dreyfus vai além. Por causa dessa irredutibilidade, uma máquina jamais poderia masterizar em jogos que exigiam algum grau de abstração como o xadrez.

Por exemplo, segundo Dreyfus, ao contrário do que supostamente se pensa na pesquisa de IA, um *grandmaster* de xadrez não pensa nas regras dos movimentos quando vai fazer uma jogada. Essas regras estão sim dentro de sua mente, mas fazem parte de um "todo", sendo incorporadas de forma "holística" em seus processos mentais. Nesse sentido, é como se as regras do xadrez estivessem "às escuras", como se fizessem parte do subconsciente, e que, devido à capacidade do jogador, pensar seria tão intuitivo que não precisaria vir a mente na forma de uma "regra" ou movimento específico. Para uma máquina, era possível aprender as regras e movimentos, mas jogar de maneira plausível a

ponto de derrotar um grande jogador era impossível. Infelizmente para Dreyfus essa crítica não durou muito. O problema é que, apesar do fato de que um enxadrista experiente realmente não precisar "pensar" no que fazer, no caso de movimentos simples, em algum momento ele *precisou* pensar. Além disso, o enxadrista, querendo ou não, para saber como fazer um movimento ou jogada x precisou aprender as regras do jogo. A partir desse caso do xadrez, e de outros argumentos desenvolvidos ao longo de *What Machines Can't Do*, os críticos de Dreyfus (Papert, 1966), (Dennett, 1984), (Russel & Norvig, 2004) argumentaram (de maneira extremamente feliz) que a critica de Dreyfus não servia à Inteligência Artificial como um todo, mas sim a uma forma específica de se programar as máquinas. Posteriormente, em seu manual de 1994, Stuart Russell e Peter Norvig comentaram esse ponto das críticas de Dreyfus e de seu irmão, Stuart Dreyfus (que foi coautor de grande parte dos livros).

> A posição que eles criticaram veio a ser chamada "Good Old-Fashioned AI" ("a boa e velha IA"), ou GOFAI, um termo cunhado pelo filósofo John Haugeland (1985). […] A GOFAI afirma que todo comportamento inteligente pode ser captado por um sistema que raciocine logicamente a partir de um conjunto de fatos e regras que descrevem o domínio. A crítica de Dreyfus não é dirigida aos computadores em si, mas a uma forma específica de programá-los. (RUSSELL & NORVIG, 2004, p. 1177)

A primeira parte da crítica de Dreyfus aparentemente foi muito bem respondida pelos cientistas da computação. Mas e a segunda? Ironicamente, essa é a parte que Dreyfus menos tratou em 1972 e ao logo da sua obra. Dissemos, no começo, que Dreyfus é um fenomenologista e que grande parte do seu trabalho se baseia na filosofia de Merleau-Ponty. Pois bem, é justamente por isso que ele considera que máquinas nunca serão capazes de ter uma mente. Convém expormos brevemente a fenomenologia e seus conceito sobre a mente, conceitos esses que são usados por Dreyfus. A fenomenologia nasceu com Bretano e Husserl no final do século XIX e desenvolveu-se ao longo da primeira metade do século XX, especialmente na França. Via de regra, a fenomenologia defende que os estados mentais são caracterizados pela consciência e pela intencionalidade, onde intencionalidade seria o "ponteiro" para onde a mente está sempre se guiando (Honderich, 1994). Seria como se, grosso modo, a mente sempre estivesse pensando em *alguma coisa,* qualquer que seja, além de ser impossível nunca pensar em nada (alguns fenomenologistas discordam desse ponto, mas não entraremos em detalhes). Mais do que isso, para Husserl, a intencionalidade seria o "motor" da consciência. Segundo descrição do mesmo, "A Intencionalidade é aquilo que caracteriza a consciência em sentido pregnante, permitindo indicar a corrente da vivência como corrente de consciência e como unidade de consciência" (Husserl, 1913).

Até recentemente, filosofia da mente e fenomenologia andaram distantes. Por exemplo, enquanto a

fenomenologia no começo do século XX já considerava o estudo das estruturas da consciência a partir do ponto de vista do individuo em primeira pessoa, a filosofia da mente analítica só veio dar especial atenção a essa questão depois da publicação do trabalho de Thomas Nagel em 1974, dando início ao estudo dos *qualia.* Apesar dessa suposta "vanguarda" da fenomenologia, algumas ponderações precisam ser feitas. Enquanto a filosofia da mente considera que a mente possuí características naturais e que essas propriedades precisam ser estudadas a partir de métodos semelhantes ao das ciências empíricas, a fenomenologia assume que, enquanto os objetos materiais são estudados a partir das suas características espaciotemporais, a mente deve ser estudada a partir de características do pensamento, isto é sensações, percepções e etc. Em resumo, a fenomenologia assume um dualismo, e alguns dos fenomenologistas mais tarde se posicionaram como anti-cognitivistas, enquanto que do lado analítico, depois de Ryle, filósofos da mente defenderam uma ontologia da mente explicitamente naturalista. Em entrevista, anos mais tarde, Dreyfus, devido a sua instrução essencialmente fenomenológica, assumiu que jamais aceitaria que a mente seria capaz de ser gerada por um meios materiais. (McCordduck, 2004)

Quando o debate em torno do trabalho de Dreyfus estava acalorado, nenhum cientista da computação, ou pesquisador da IA, respondeu diretamente a questão da intencionalidade e era pouco provável que pudessem obter algum sucesso nesse tipo de discussão. Como discutido no capítulo anterior, para se aceitar que um máquina é capaz de ter mente é preciso, a principio, assumir uma posição ontologicamente materialista, coisa que Dreyfus jamais assumiria. Nesse sentido, a discussão entre Dreyfus e a pesquisa em IA parecia ser um choque de mundos distintos, onde Dreyfus parecia estar munido de argumentos extremamente fortes, porém, infelizmente grande parte do seu trabalho não foi levado a sério. Ao invés de adotar uma posição mais "fenomenológica" (que provavelmente teria sido seu maior trunfo) Dreyfus resolveu fazer valer seu tempo como consultor na RAND e criticar a Inteligência Artificial a partir de aspectos técnicos e se valendo de comparações diretas com a mente humana. Além disso, muito do trabalho de Dreyfus, especialmente as primeiras publicações, possuem tom polemista, com o filosofo fazendo acusação quanto à incapacidade de uma área que tinha começado formalmente a pouco menos de vinte anos. O resultado foi que muitas das coisas que Dreyfus se concentrou em dizer que a Inteligência Artificial não alcançaria como, por exemplo, aprendizado, tomada de decisões e etc, estão agora incorporadas à pesquisa de Inteligência Artificial, o que parece refutar sua posição de que a Inteligência Artificial é impossível.

3.1 JOHN SEARLE: A INTENCIONALIDADE LEVADA ALÉM

Se Dreyfus tratou mais dos aspectos técnicos e menos da intencionalidade no seu trabalho contra a Inteligência Artificial, o mesmo não pode ser dito de John Searle. Searle é um filosofo que segue uma linha de pesquisa muito diferente da de Dreyfus. Desde o começo de sua carreira interessou-se profundamente pela linguagem e em 1969 lançou seu primeiro livro sobre o tema dos atos de fala (chamado de *Speech Acts: An Essay in the Philosophy of Language*), teoria desenvolvida inicialmente por J. L. Austin e que foi, posteriormente, expandida por Searle e outros pesquisadores. Seu trabalho na filosofia da mente, assim como na filosofia da linguagem, é indispensável, e a sua critica à ideia de máquinas com mentes é provavelmente a mais influente que existe. Quando estudamos a crítica de Searle, percebemos que existem duas formas de fazê-lo: primeiro é estudá-la separada do seu trabalho, concentrando-se apenas nos ensaios e artigos que o acadêmico publicou cronologicamente; a segunda é partir de um olhar sobre a sua obra, o que ajuda a esclarecer porque Searle toma certos argumentos e quais as suas razões para isso. Não pretendemos fazer deste capítulo um dossiê sobre o trabalho de Searle, mas um ou outro fragmento de algumas das suas obras tomaremos emprestado para deixar a discussão mais rica.

A crítica de Searle contra a Inteligência Artificial aparece de formas diferentes ao longo do tempo, ainda que estejam ligadas entre si. Ao contrário de Dreyfus, Searle usa a distinção entre Inteligência Artificial “Fraca” e Inteligência Artificial “Forte”, e não tem interesse em atacar a Inteligência Artificial como um todo. Ao contrário, ele acredita que o campo é uma valiosa forma de se entender aspectos da mente humana. Seu problema está com o conceito de Inteligência Artificial Geral ou “Forte” como ele mesmo a chama. Dos artigos e livros que ele escreveu sobre o tema, quatro se destacam, sendo o primeiro *Minds, Brains, and Programs*, publicado em 1980, o segundo *Minds, Brains, and Science*, de 1984. O terceiro e provavelmente mais famoso, *Is the Brain's Mind a Computer Program?* de 1990, publicado na *Scientific American*. E o quarto e último *The Mystery of Consciousness*, de 1997. Ele também viria a fazer alguns comentários breves sobre o tema posteriormente, como por exemplo, um artigo para o *The Wall Street Jornal* de 2011, sobre o supercomputador da IBM, o Watson. Nos interessa aqui apenas os artigos de 1980 e 1990. Começaremos pelo de 1980, o segundo veremos apenas de maneira breve e voltaremos a falar dele especialmente no quarto capítulo.

Começamos esse trabalho falando do ensaio *Computer Machinery and Intelligence* de Alan Turing. Discutimos o que Turing pretendia e chegamos à conclusão de que seu experimento é inteiramente behaviorista e sua máquina inteiramente simbólica. No artigo de 1980, Searle tinha como alvo

justamente essas duas questões, não exatamente o Teste de Turing, mas sim programas de perguntas e respostas (chamados de *chatterbots*) que pretendessem passar no teste, como o programa desenvolvido pelos pesquisadores Roger Schank e Peter Alberson em 1977, bem como o programa ELIZA de Weizenbaum (Weizenbaum 1967). Para ele era improvável, senão impossível, concluir que uma máquina inteiramente simbólica fosse capaz de pensar baseando-se apenas em um teste comportamental. Em razão disso, Searle criou um experimento mental para ilustrar a sua teoria que ficou conhecido como *The Chinese Room*. Parafraseando Searle, o argumento funciona da seguinte forma:

> Suponha que eu esteja trancado em um quarto e receba um grande lote de frases em chinês. Suponha, além disso, que eu não conheça chinês, seja escrito ou falado, e que nem mesmo tenha certeza de que poderia reconhecer a escrita chinesa como uma escrita chinesa distinta, digamos, da escrita japonesa ou dos rabiscos sem sentido. Para mim, a escrita chinesa não passa de rabiscos sem sentido. Agora, suponha ainda que, após esse primeiro lote de frases em chinês, eu receba um segundo lote junto com um conjunto de regras para correlacionar o segundo lote com o primeiro. As regras estão em inglês e eu entendo essas regras tão bem quanto qualquer outro falante nativo de inglês. […] Do ponto de vista externo – do ponto de vista de alguém lendo minhas "respostas" - as respostas para as questões chinesas e as questões inglesas são igualmente boas. Mas no caso chinês, ao contrário do caso inglês, produzo as respostas manipulando símbolos formais não interpretados. No que diz respeito aos chineses, simplesmente me comporto como um computador; Realizo operações computacionais em elementos formalmente especificados. Para os propósitos dos chineses, sou simplesmente uma instanciação do programa de computador. (SEARLE, 1980, p. 418)

Ao longo da sua crítica, Searle vai readaptar o experimento de acordo com a situação, de forma que resumos simplistas como "*The Chinese Room* é como um movimento mecânico de usar um dicionário para traduzir sentenças do Chinês para o inglês e do inglês para o chinês" também são válidos para entender a lógica do teste. O importante, segundo Searle, não é o experimento propriamente dito, mas sim a teoria por trás dele, que no caso é a teoria de que manipular símbolos não produz ou necessita de qualquer forma de pensamento. No artigo de 1980, Searle tem como alvo uma teoria específica de Inteligência Artificial Forte, que no caso é aquela que foi chamada de "GOFAI", citada acima quando falamos da crítica de Dreyfus. Além disso, ele também critica (de maneira indireta) a Teoria Computacionalista Clássica e o Funcionalismo. O grande trunfo de Searle, diferente de Dreyfus, foi ter utilizado o argumento da intencionalidade.

Ainda que o conceito básico de intencionalidade em Searle seja o mesmo que o de Dreyfus e dos fenomenologistas, a intencionalidade de Searle, em suas nuances, não é nem de perto a mesma intencionalidade discutida pelos continentais. Primeiro, ela é puramente materialista e se baseia em aspectos causais da mente, segundo há uma forte relação entre a intencionalidade e os Atos de fala

(Searle, 1983). Para Searle, o cérebro é uma máquina, de forma que no decorrer do artigo de 1980 ele vai explicar que a frase “Máquinas podem pensar” não é necessariamente incorreta, desde que você esteja considerando uma máquina biológica como o cérebro. Para entender um pouco da relação entre cérebro e intencionalidade, vale expor o conceito do *naturalismo biológico*, que é a ideia “mestre” que guia o artigo de 1980. Searle propôs essa ideia pela primeira vez em 1980, onde o naturalismo biológico é uma teoria sobre a relação entre consciência e corpo (ou seja, cérebro) e, portanto, sua abordagem ao problema mente-corpo que pode ser definida por duas teses principais: 1. Todos os fenômenos mentais, desde dores, cócegas e coceiras até os pensamentos mais abstratos, são causados por processos cerebrais de nível inferior; 2. Os fenômenos mentais são características de nível superior do cérebro. Isso implica que o cérebro tem os poderes causais certos para produzir intencionalidade; (Searle, 2004).

Com o naturalismo biológico, Searle propõe que “esqueçamos” o tradicional problema mente-corpo como definido ao longo da história da filosofia e nos concentremos apenas na parte “biológica”, ou seja, o cérebro e aquilo que percebemos. Enquanto que a história da filosofia costuma considerar a consciência como aspecto separado, o naturalismo biológico argumenta que ela é parte do processo evolutivo animal, assim como são os olhos e a visão. Na perspectiva de Searle, embora seja razoável considerar a consciência como um produto do objeto natural chamado “cérebro”, não é razoável, estudar a mente com as mesmas ferramentas das ciências naturais porque a mente não é um objeto “visível” sobre nenhum aspecto físico comum da ciência. Em miúdos, a única maneira de compreendermos a consciência, segundo Searle, é por meio da linguagem, uma vez que ela é a única que pode revelar os atos intencionais dos estados mentais dos indivíduos (Searle, 1983).

Não é fácil compreender como a filosofia da linguagem de Searle e como ela está ligada com a sua crítica a máquina com mentes. No entanto, no artigo de 1980, ele afirma que máquinas não podem ter mentes porque não tem intencionalidade, e que a intencionalidade é um estado causal do cérebro. Para Searle, materialidade importa, e o fato das máquinas instanciarem um programa é um dos problemas da Inteligência artificial forte, uma vez que o cérebro não instância a mente como um programa de computador, mas a mente é causada pelo cérebro, uma reação existente graça a seus processos biofísicos. Em nenhum momento, no artigo de 1980, Searle cita diretamente seu naturalismo biológico, no entanto, a questão fica subentendida, especialmente quando ele se ocupa em responder as objeções que recebeu quanto ao seu argumento. Essa objeções (que são expostas no final do artigo e que Searle responde-as individualmente) apareceram ao longo do tempo em que ele apresentou palestras sobre a questão das máquinas com mentes, antes de condensar sua ideia no

artigo de 1980. Elas são seis, vamos apresentá-las em tópicos retirando-as diretamente do artigo original, ou seja, da mesma maneira que Searle as apresentou, porém, só iremos comentar de maneira mais detalhada as respostas que Searle considerou mais relevantes e que, vez ou outra, reapareceram no debate gerado pelo seu trabalho. Eis um resumo das objeções:

1. *"The Systems reply"*

"Embora seja verdade que o indivíduo que está trancado na sala não entende a estória, o fato é que ele é apenas parte de um todo, o sistema, e o sistema entende a estória. (No caso, o quarto entende Chinês)"

2. *"The Robot reply"*

"[…] Suponha que coloquemos um computador dentro de um robô, e este computador não apenas receba símbolos formais como entrada e distribua símbolos formais como saída, mas sim opera de tal maneira que faça algo parecido com "sentir", andando, movendo-se, martelando pregos, comendo, bebendo—qualquer coisa que você queira. O robô, por exemplo, pode ter uma câmera acoplada que o permita "ver", teria braços e pernas que lhe permitiriam "agir", e isso seria controlado pelo cérebro de seu computador."

3. *"The Brain simulator reply"*

"Suponha que projetemos um programa que não representa as informações que temos sobre o mundo, como as informações nos *scripts* de Schank, mas simula a sequência real de disparos dos neurônios nas sinapses do cérebro de um chinês nativo quando ele entende estórias em chinês e dá respostas para eles. Nesse caso há entendimento."

4. *"The Combination reply"*

"Imagine um robô com um computador em forma de cérebro alojado em sua cavidade craniana, imagine o computador programado com todas as sinapses de um cérebro humano, imagine que todo o comportamento do robô é indistinguível do comportamento humano, e agora pense na coisa toda como um sistema unificado e não apenas como um computador com entradas e saídas. Certamente, em tal caso, teríamos que atribuir intencionalidade ao sistema."

5. *"The Other minds reply"*

"Como você sabe que outras pessoas entendem chinês ou qualquer outra coisa? Apenas pelo seu

comportamento. Agora o computador pode passar pelos testes tão bem quanto nós podemos (em princípio), então se você vai atribuir cognição a outras pessoas, você deve, em princípio, também atribuí-lo a computadores."

6. *"The Many mansions reply"*

"Todo o seu argumento pressupõe que a IA é apenas analógica e os computadores digitais. Mas isso é apenas o presente estado da tecnologia. Quaisquer que sejam esses processos causais que você diz que são essenciais para a intencionalidade (assumindo que você está certo), eventualmente poderemos construir dispositivos que tenham esses processos causais, e isso será inteligência artificial."

Comecemos pelo *Systems Reply.* Para responder a essa questão Searle assume que poderíamos fazer com que a pessoa dentro do quarto "mentalizasse" o experimento *The Chinese Room* em sua cabeça. Nesse sentido ele meio que assume novamente que o experimento em si não é a parte mais importante, mas sim a parte teórica que se deriva. Em um exercício lúdico e aproveitando os avanços que temos hoje na tecnologia, em comparação a 1980, seria mais ou menos como se, ao conversar pela internet com um falante de Chinês, um usuário falante de língua inglesa apanhasse todas as sentenças em chinês e as jogassem num tradutor. A única coisa que o usuário fez foi um movimento "mecânico", por assim dizer, de "traduzir" as frases e fornecer respostas corretas. Entretanto, a maneira como Searle trata esse questionamento é um tanto quanto peculiar. Ele usa um tom mais polemista ao dizer no artigo que apenas "partidários da ideologia da IA forte seriam capazes de um argumento como esse", ao mesmo tempo que ele faz, novamente, uma espécie de crítica ao funcionalismo enquanto busca investigar quais as possíveis origens de um argumento como esse. O caso é que, segundo Searle, se trata de uma resposta que leva em conta a razão "funcional" da mente, isto é, o sistema inteiro em funcionamento cria uma "consciência" a respeito daquilo que se passa entre receber as sentenças e respondê-las. Ainda que não cite explicitamente o funcionalismo, a descrição é muito parecida com o modo que o funcionalismo compreende a questão da consciência. A resposta de Searle nesse caso é que esse estado "funcional" é falho porque, se por ventura a mente entende chinês apenas pela manipulação funcional dos símbolos, outro órgão como o estômago, por exemplo, também exerce uma atividade "funcional" (como a digestão) e mesmo assim nós não estamos nem de longe "conscientes" sobre o que acontece no estômago. O questionamento *System Reply* retorna de maneira diferente ou ideias derivadas dele reaparecem nos próximos debates que Searle vem a ter ao redor da questão das máquinas com mentes, especialmente em 1990.

A resposta que Searle concede pra o caso do *Robot Reply* é curta, mas é algo que tanto um pesquisador de IA quanto um filósofo podem concluir sem necessariamente estarem alinhados com ele. Ainda que a resposta assuma que não basta a manipulação de símbolos linguísticos para que haja "entendimento", ela não deixa clara de que forma o robô manipularia os outros "sentidos" que ele viria a receber, de modo que se subentende que ele o faria da mesma forma que o fez com as sentenças linguísticas. Imagens, cheiros, sons, tudo isso seria transformado em símbolos para o Robô. Junto com o *Systems Reply*, essa resposta voltaria ao debate, especialmente, quando considerarmos redes conexionistas em que existe um processamento de imagens, sons e etc, e de uma maneira diferente da AI simbólica tradicional. Nós expandiremos esse debate no terceiro e quarto capítulo.

Vamos adiantar as respostas do *Other minds* e *Many Mansions* antes de prosseguirmos para as respostas do terceiro e quarto argumento, já que teoricamente esses são os argumentos mais simplórios. O *Other Minds Reply* é um tanto quanto parecido com a objeção que Turing debateu em *Computing Machinery and Intelligence*. Enquanto Turing responde que considerar o problema *Other Minds* é uma atitude solipsista (um tanto quanto erroneamente) Searle responde que o principio da ciência cognitiva, diferente do behaviorismo duro, é considerar que existem estados mentais tal como a física considera que existe gravidade e outras forças invisíveis, ainda que, diferente da física, não exista consenso a respeito da descrição desses estados mentais. *Many Mansions* é mais uma previsão ou um ponto de vista do que um contraponto. Apesar de não sermos capazes de atribuir intencionalidade a IA nesse momento, talvez sejamos capazes num futuro próximo. Searle não nega que talvez isso seja possível, bem como ele não nega que máquinas artificiais talvez sejam capazes de pensar no futuro, o seu problema é com um projeto específico de artificialidade, o projeto que argumento que a mera instanciação de um programa que manipula símbolos é o bastante para produzir uma mente.

O *Brain Simulator Reply* encontra uma posição peculiar na literatura. Ele poucas vezes é citado diretamente, e nunca é citado como a resposta mais influente, apesar de se tratar de um possível argumento conexionista. Em nenhum momento os termos "redes neurais" ou "conexionismo" são citados diretamente no artigo, nem pela objeção, nem por Searle. Já existia nessa época uma pesquisa em redes neurais artificiais, mas o campo ainda não tinha uma formalização como ocorreu em 1986. Fato é que, de 1984 em diante, muitas das objeções que Searle vai receber contra o seu experimento partem justamente da ideia de processamento paralelo de informações. A resposta que ele dá para essa "primeira objeção conexionista" em 1980 é um tanto quanto desajeitada. Primeiro

ele considera que essa resposta é um tanto quanto estanha para um partidário da IA forte. Sob certo aspecto, é mesmo. Nos anos 1970, o debate dentro da teoria computacional assumia principalmente que a manipulação simbólica era suficiente para criar representação, essa posição só perdeu espaço com o desenvolvimento da pesquisa em processamento paralelo distribuído, anos mais tarde. Ademais, Searle está convicto de que a maioria dos partidários da IA forte consideram que apenas a manipulação simbólica é capaz de, não apenas criar entendimento, mas também de construir algum programa inteligente plausível. Nesse sentido ele comete em um erro semelhante ao de Dreyfus em 1972, uma vez que já existia uma pesquisa em processamento paralelo na época (Minsky, Papert, 1969), ainda que, a favor de Searle, essa pesquisa estivesse muito longe de ser popular nos círculos acadêmicos. Seu contra-argumento é parte de um proposta semelhante àquele do *Systems Reply*, onde ele adapta o experimento e transforma a pessoa dentro do quarto num "operário" que puxa e empurra alavancas num movimento semelhante ao dos neurônios artificiais. Ao invés de apanhar as fichas com os textos em chinês e procurar uma tradução para elas, o indivíduo dentro do quarto manipularia um jogo de chaves e tubos que sintetizariam as principais funções de um neurônio. Em resumo, diz Searle, ainda que uma rede neural artificial simule o funcionamento dos neurônios do cérebro, o que é algo mais promissor do que a solução puramente simbólica, ela não é suficiente para produzir intencionalidade, uma vez que não artificializa as partes corretas da rede neural biológica que produzindo intencionalidade. Em resumo, para Searle, mesmo que uma rede neural artificial seja um modelo tão perfeito da rede biológica ela não é capaz de "captar" tudo aquilo que o cérebro biológico é capaz. Nesse caso, o chinês com redes neurais artificiais na cabeça até poderia ter um comportamento semelhante ao de um chinês de verdade, mas careceria de entendimento.

> Enquanto simular apenas a estrutura formal da sequência de disparos de neurônios nas sinapses, não terá simulado o que importa no cérebro, ou seja, suas propriedades causais, sua capacidade de produzir estados intencionais. (SEARLE, 1980, p. 421)

Quanto ao *Combination Reply*, onde as três resposta são combinadas (um cérebro com redes neurais artificiais dentro de um robô que é capaz de perceber e interagir com o seu mundo, onde não apenas os "neurônios" individuais do robô entendem aquilo que ele percebe, mas todo o seu cérebro), uma vez que Searle, segundo suas próprias perspectivas, já demonstrou que os outros três argumentos são incorretos, não faria sentido juntar os três na expectativa de que eles produzissem um resultado diferente. A única coisa que Searle concorda, nesse caso, é que de fato seria tentador atribuir entendiamento a um robô desse tipo. Posteriormente, Searle não retornaria com o argumento da pessoa puxando canos dentro de um quarto, ainda que insistisse com o experimento. Vamos guardar um pouco dessa discussão quanto às redes neurais para o próximo capítulo onde trataremos dessa

linha de pesquisa e como ela concedeu um “segundo round” ao debate sobre máquinas com mentes. Os argumentos que veremos, especialmente os dos Churchlands, partem, justamente, desse ponto de vista. Querendo ou não, em meados dos anos 1990, poucos pesquisadores e adeptos da inteligência artificial forte estavam dispostos a defender a IA Simbólica. Nesse sentido, Churchland (1990) dá razão a Searle.

> A manipulação de símbolos governados por regras não é seu modo básico de operação. O argumento de Searle é dirigido contra as máquinas *SM* governadas por regras [...] nós, e Searle, rejeitamos o teste de Turing como uma condição suficiente para a inteligência consciente. Em certo nível, nossas razões para fazer isso são semelhantes: concordamos que também é muito importante como a função de *input-output* é alcançada; é importante que as coisas certas estejam acontecendo dentro da máquina artificial; (CHURCHLAND, 1990, p. 37)

4. A RESPOSTA CONEXIONISTA

Conexionismo é um movimento das ciências cognitivas que espera explicar as habilidades intelectuais humanas a partir do funcionamento das redes neurais artificiais. Paul M. Churchland e Patrícia S. Churchland foram os primeiros filósofos a utilizarem o conexionismo como um argumento favorável à Inteligência Artificial Forte (Churchland, 1990), Isso porque as redes neurais artificiais são um modelo substancial das redes neurais biológicas, que pretendem simular as sinapses humanas. Resumidamente, essas redes são compostas de unidades que são modelos matemáticos substanciais dos neurônios do cérebro. O Interesse da filosofia nas redes neurais artificiais ocorre tanto pela suposta possibilidade de representação mental que elas podem possuir, isto é, a capacidade de existir algum "pensamento", quanto ao debate a respeito da teoria computacional, em que a mente seria um processador de informações. Tanto o computacionalismo clássico como o conexionismo argumentam que a mente processa informações, a diferença está que enquanto o computacionalismo clássico defende uma via por meio do processamento simbólico, o conexionismo argumenta em defesa da excitação de sinapses múltiplas que "guardariam" elementos da percepção. Nosso interesse aqui é mais limitado do que a discussão geral que envolve o conexionismo, queremos saber apenas como essas redes funcionam e como elas poderiam ser usadas em favor da Inteligência Artificial Forte.

Em resumo, uma rede neural artificial consiste em unidades (ou neurônios) agrupados num padrão de conexão. Na figura abaixo podemos ver um exemplo de um neurônio artificial.

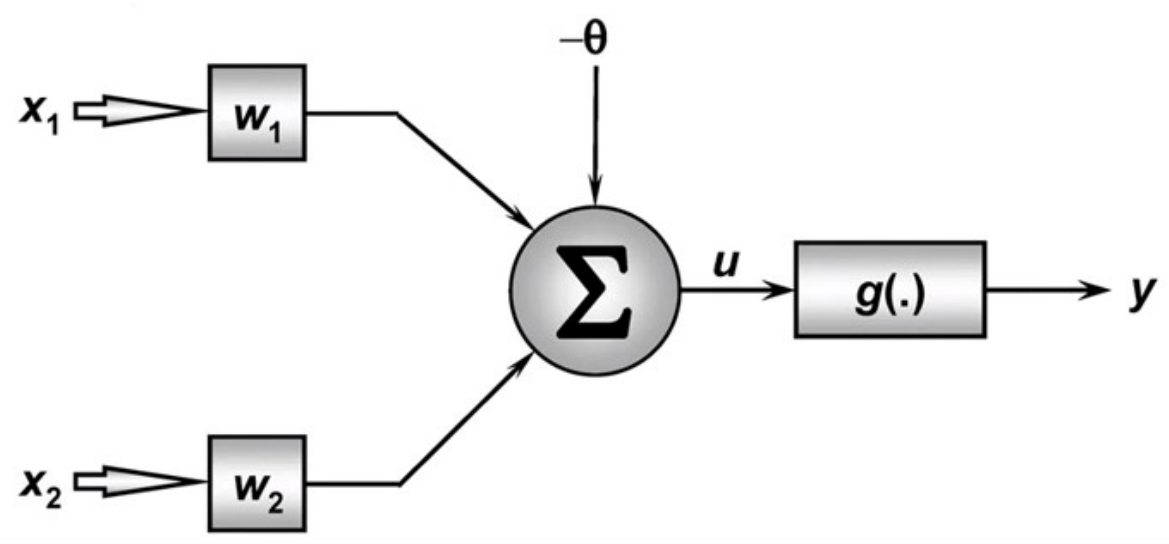

Figura 2: Neurônio do tipo perceptron. X¹ e X² representam as entradas do neurônio, g() representa a função de ativação, isto é, a função que determina se o neurônio deve ser ativo ou

não, y representa a saída do neurônio. W^1 e W^2 representam os pesos das entradas do neurônio, enquanto -Θ representa o Bias, que serve pra normalizar o somatório (Σ) do valor das entradas. U representa a saída do somatório.

Numa rede neural, as unidades em geral são separadas em três classes unidas entre si por meio das suas saídas. Unidades de entrada (*input units*), que recebem as informações; unidades de saída, (*output units*) por onde saem os resultados e unidades escondidas (*hidden units*), que ficam entre as *input* e *output units*. Um exemplo análogo bastante simples para entender o funcionamento dessas redes é, enquanto as *input units* seriam, por exemplo, os nervos sensoriais, como o tato, por exemplo, que entra em contato diretamente com os objetos, as *hidden units* seriam os neurônios de processamento desses sentidos, seriam como os neurônios que ficam no cérebro. Enquanto que os *output units* seriam os neurônios de resposta, ou seja, os "nerônios motores". Cada unidade do tipo *input* possui um valor de ativação que representa alguma caraterística externa à rede. Uma unidade *input* envia o seu valor de ativação para cada uma das unidades escondidas com a qual está conectada, cada uma dessas unidades calcula seu próprio valor de ativação dependendo do valor de ativação que recebe das unidades de entrada. O sinal então é passado das unidades de saída. O padrão de ativação configurado por uma rede é determinado pelos pesos ou pela força da conexão entre as unidades. O valor desses pesos pode ser positivo ou negativo, um peso negativo representa a inibição de atividade pela unidade que envia. O valor de ativação para cada unidade recebedora é calculado de acordo com uma simples função de ativação. Essas funções variam em detalhes, mas o resumo geral é que a função de ativação faz um "somatório" de todos os sinal de recebeu das outras unidades.

Tendo esse funcionamento em vista, o conexionismo presume que a cognição pode ser explicada por redes neurais que operam dessa maneira, uma vez que se assume que todas as unidades de uma rede calculam a partir de uma função de ativação em comum, a atividade intelectual humana deve depender, tal como nas redes artificiais, primariamente das "configurações dos pesos" entre as unidades.

Apesar das redes neurais terem ingressado no debate publico nas últimas décadas, sua história é relativamente antiga. O primeiro conceito de Redes Neurais surgiu em 1943, a partir do trabalho de Warren McCulloh e Walter Pitts, que compuseram um modelo matemático de um neurônico biológico. Grande parte da noção que McCulloh e Pitts queriam sintetizar era o caráter "*all-or-nothing*" de um neurônio biológico. (1943). Por causa desse trabalho, ambos são considerados por muitos como os pais da inteligência artificial, mas como já vimos, esse tópico é questionável. O

primeiro método de treinamento de uma rede neural surgiu em 1949, e foi chamado de "*Hebb's rule*". (Hebb 1949). Entre os modelos de neurônio mais conhecidos está o perceptron, usado até hoje, que foi desenvolvido por Frank Rosenblatt (Rosenblatt 1958). É a partir do perceptron que as redes neurais artificiais tornam-se mais plausíveis e o interesse acadêmico por elas aumenta, muito graças à capacidade deste neurônio em reconhecer padrões simples. No entanto, até 1980 as redes neurais vão ocupar um espaço secundário na pesquisa de inteligência artificial, em razão do baixo poder de processamento dos computadores daquela época. Por exemplo, sessões de treinamento de uma rede neural com algumas poucas unidades chegavam a levar horas ou dias. Além disso, o trabalho de Marvin Minsky e Seymor Papert (1969), ambos fundadores da pesquisa em inteligência artificial, e fortemente envolvidos no debate da IA Forte, colocou alguns obstáculos para pesquisa em redes neurais. Em seu livro *Perceptrons: An Introduction to Computation Geometry*, os autores discutem as limitações das redes neurais, entre as quais, estava a incapacidade de divisão de padrões que não fossem linearmente separáveis. Esse trabalho foi um golpe duro para a pesquisa em redes neurais, que só veria seu interesse reflorescer nos anos 1980. O ressurgimento das redes neurais está intimamente ligado ao conexionismo, muito devido ao trabalho de David Rumelhart (1986), em seus dois volumes, chamados *Parallel Distributed Processing*. Hoje, “Parellel Distributed Processing” e “Conexionismo” são termos que, devido a toda discussão cognitiva, significam a mesma coisa. Nesse trabalho, Rumelhart e seu colaborador, James McClelland, afirmam que a cognição humana funciona a partir de um processo de processamento paralelo de informações. Fortemente inspirados nesse trabalho, Patricia Churchland e Paul M. Churchaland vão publicar em 1990 uma “resposta” ao experimento *The Chinese Room* de John Searle, além de, entre outras coisas, introduzir o conexionismo no debate da Inteligência Artificial Forte.

Paul Churchland e Patricia Chruchland são dois filósofos da mente. Apesar de possuírem trabalhos assinados individualmente, os dois cooperam em grande parte dos seus trabalhos. Por causa disso é relativamente difícil discernir onde exatamente começa e acaba o pensamento de Paul Churchland e onde começa e onde acaba o pensamento de Patricia Churchland. Muitos comentadores, a título de facilitação, referem-se a ambos apenas como “Os Churchlands” mesmo em trabalhos que foram assinados apenas por um ou por outro. É isso o que faremos aqui. Assim como Searle baseia sua crítica à IA Forte em seu trabalho linguístico, a defesa da possibilidade eventual de máquinas possuírem mentes no caso dos Churchlands reside em um poderoso *background* trabalhado pelos dois ao longo de uma década de produção. Comentaremos esse trabalho conceitualmente. Nenhuma figura na filosofia colocou as redes conexionistas em maior evidência do que eles. Em certo sentido,

é possível dizer que ambos tem um papel fundamental no sucesso desse campo, isso porque os Churchlands aplicaram o conexionismo a varias áreas da filosofia, desde a filosofia da mente até a ética. O sucesso deles em trazer uma perspectiva experimental para uma área em que a especulação é, sob certo ponto de vista, fundamental, reside na capacidade que ambos tem de investigar as questões filosóficas dentro da neurociência e apontar possíveis soluções para esses problemas. Nesse sentido, o naturalismo de ambos, se for possível colocar em níveis, está em uma escala maior do que o de Searle. Quine afirmou que filosofia e ciência são contínuos, os Churchlands concordam plenamente.

O primeiro grande trabalho dos Churchlands no conexionismo foi publicado em 1986, sob o título "*Some reductive strategies in cognitive neurobiology"*. A inspiração começou a partir do trabalho de Andras Pellionisz e Rodolfo Llinás. Pellionisz e Llinás (1979) argumentaram que as funções do cérebro são representadas em espaços multidimensionais, que as redes neurais devem, portanto, ser tratadas como "objetos geométricos" e que a "linguagem interna do cérebro é vetorial". Ambos também argumentaram em favor das redes neurais quando comparadas a estrutura computacional clássica de Von Neumann, afirmando que as redes neurais são mais resistentes a eventuais erros em sua atividade de processar informações. Tanto a questão da função cerebral, quanto as críticas ao computacionalismo clássico, são temas caros aos Churchlands, aos quais ambos beberam diretamente de Pellionisz e Llinás.

Com sua perspectiva naturalista, bem como sua capacidade de trabalhar com resultados experimentais da neurociência, os Churchlands apresentam suas teorias de maneira simpática aos filósofos em geral, mas atacam concepções clássicas de maneira vigorosa. Ambos afirmam que a epistemologia deve ser naturalizada, sobretudo as discussões epistemológicas sobre questões cognitivas, como é o caso da percepção, dos sistemas mnemônicos e dos sistemas inferenciais, que são temas caros à pesquisa neurocientífica experimental nas últimas décadas. Perguntar sobre como se dá o conhecimento é perguntar sobre como o cérebro processa conhecimento (Churchland, 2002). Além da epistemologia, a Psicologia também é alvo dos Churchlands, e de uma maneira até mais vigorosa e polêmica. Alguns trabalhos mais recentes ambos assinam em coautoria com os neurocientistas Pellionisz e Llinás, como *The Mind-Brain Continuum* (1996), onde os Churchlands buscam investigar os processos psicológicos a partir de uma perspectiva neurocientífica. Para eles a psicologia comum comete muitos erros ao descrever o comportamento e a vida mental humana, e sendo assim, seria interessante uma profunda alteração ou talvez mesmo uma eliminação em razão

da neurociência (Churchland, 1992). O que fica claro logo de início é que, para eles, a pesquisa científica deve encontrar todas as respostas. Se ainda não encontrou, logo as encontrará. Esse tópico é tema comum em grande parte do trabalho dos Churchlands. Se o livro *Intentionality* (1983) de John Searle serve como o pilar para a sua crítica quanto à questão das máquinas com mentes, algo semelhante acontece com o artigo de 1986 dos Churchlands, onde resumidamente eles vão "trazer" para a filosofia a discussão sobre a questão da representação em sistemas computacionais paralelamente distribuídos. Seus argumentos, grosso modo, são uma defesa de uma estrutura redutiva explicativa de como os processos cognitivos funcionam através dos modelos artificiais das redes neurais artificiais.

O artigo de 1990, *Could a Machine Think?*, publicado na *Scientific American,* sumariza centenas de páginas de volumes de um argumento favorável a possibilidade de máquinas com mentes. A estratégia argumentativa é simples: Mostrar como uma rede conexionista pode ser um meio possível para a consciência. Logo no subtítulo do artigo, os Churchlands concordam a respeito da impossibilidade da Inteligência Artificial Clássica ser incapaz de ter algum tipo de consciência. Observando o trabalho desenvolvido por eles é possível assumir que não se trata de um movimento doloroso para eles assumir tal posição. Enquanto os argumentos funcionalistas a favor da IA simbólica delineiam que, grosso modo, nosso cérebro processa suas representações por meio de predicados linguísticos e que, desse modo, programas que operam no degrau simbólico podem ter algum nível mental, os Churchlands não concordam com essa afirmação. Para eles, as representações mentais ocorrem em um nível associacionista, semelhante a aquele esquematizado por Hume no século XVIII, onde, ao invés de predicados linguísticos, a mente processa "imagens". Mesmo assim, na primeira parte do artigo, os Churchlands vão reavaliar as críticas a IA Clássica feitas por Searle e elaborar, a partir do ponto de vista deles mesmos, uma possível defesa para esses argumentos. O motivo para isso é simples, conforme discutimos, Searle tinha em mente os *Chatterbots* quando escreveu sua crítica original em 1980, mas para ele sua crítica não vale apenas para um tipo específico de Inteligência Artificial. O *The Chinese Room* lida não apenas com IA Simbólica, mas com todas as formas de IA, sejam elas feitas de regras ou por meio de processamento paralelo distribuído. Como os Churchlands descrevem, para Searle, a Inteligência Artificial não é mais do que um *mockup* vazio da nossa própria mente. Ainda que no artigo de 1980 houvesse apenas um esboço da sua crítica às redes neurais, pouco tempo depois Searle refinaria essas críticas, e já no artigo de 1990 (publicado na mesma edição que o dos Churchlands) há um

questionamento muito ácido quanto a essas redes conexionistas. Essas críticas serão examinadas no próximo capítulo que concluí essa discussão.

A pergunta “pode uma máquina pensar?” pode ser interpretada como: “Pode uma máquina que manipula símbolos semelhantes às regras da estrutura sensível pensar?”. Os Churchlands resumem o raciocínio favorável em dois pontos. O raciocínio não é necessariamente original e baseia-se na tradição comportamental da filosofia da mente. Em suma: 1. Cada função efetiva pode ser recursivamente computável. Isto é, cada procedimento possível é uma “rota” que pode ser sumarizada por regras (Quando beliscado, diz-se “ai”, por exemplo). Portanto, pode-se construir um programa que consiga buscar qual a melhor “rota” a ser seguida dependendo da situação em que se encontra. 2. Qualquer uma dessas funções efetivas podem ser, conforme demonstrou Turing, computadas por uma máquina que manipule símbolos. Nesse sentido, qualquer máquina de estados discretos (computador) pode ser programada para agir de maneira inteligente. Vimos esses dois argumentos ainda no primeiro capítulo, bem como as respostas de Dreyfus e Searle para eles. Nesse sentido, é interessante destacar a tréplica que os Churchlands dão a Dreyfus e, em especial, a Searle. Em sua crítica, um dos principais aspectos que tanto Searle quanto Dreyfus vão citar de maneira desfavorável à IA Simbólica é sua ausência de intencionalidade. Mais tarde, no argumento de Searle, essa ausência de intencionalidade se transforma em ausência semântica. Mas existe um problema que Searle não pontua em sua crítica, que é como aconteceria o processo de dar significado aos objetos da percepção. Nesse sentido, existe uma lacuna explicativa que nem mesmo o Searle dá conta de resolver, embora ele acuse a IA Simbólica de não possuir uma explicação nesse ponto. Nesse sentido, os Churchlands argumentam que uma réplica viável da pesquisa em IA Simbólica a Searle poderia ser algo como: Os pesquisadores se baseiam naquilo que eles já sabem e não podem ser culpados por uma suposta ignorância que nem mesmo os filósofos da mente dão conta de responder. Turing, de uma maneira ou de outra, respondeu algo semelhante em seu artigo de 1950. Sua inteligência artificial se baseava naquilo que já sabíamos que poderia ser uma comprovação de alguma possível atividade cognitiva.

Em todo caso, os Churchlands não estão a favor da IA Simbólica. Apesar de reconhecerem seu potencial, por mais que uma IA Clássica consiga jogar xadrez ou manter uma conversa simples, é improvável que faça mais do que isso. Antes de explicar diretamente as redes conexionistas e a possibilidade de consciência por parte delas, os Churchlands se concentram em desmantelar o

argumento de Searle no *The Chinese Room*. Após o artigo de 1980, Searle resumiu a ideia por trás do *Chinese Room* em três axiomas e uma conclusão. São eles:

1. Primeiro axioma: Os programas de computador são formais (sintáticos).

2. Segundo axioma: As mentes humanas têm conteúdos mentais (semântica).

3. Terceiro axioma: A sintaxe por si só não é constitutiva nem suficiente para a semântica.

4. Conclusão: Os programas não são constitutivos nem suficientes para as mentes.

Esses três axiomas não foram definidos por Searle em 1980, mas apenas em 1984, quando ele vai esquecer a intencionalidade e vai se concentrar no problema do significado, os motivos dessa virada, como já vimos no capítulo anterior, estão escorados na sua filosofia da linguagem. Ademais, os Churchlands afirmam que dos três axiomas e sua conclusão, o terceiro axioma é o que carrega noventa por cento do argumento de Searle. O *Chinese Room*, eles concordam com Searle, é um experimento meramente ilustrativo. Não importam críticas como: "O experimento é ridiculamente lento" ou "não é exatamente assim que um programa funciona". Para ambos, o erro do *Chinese Room*, ou mais especificamente dos axiomas que derivam dele, é um erro lógico. Para demonstrar isso eles constroem um experimento imaginário com um senso imaginativo semelhante ao de Searle, usando as descobertas de James Maxwell no séc. XIX. O objetivo é mostrar, em um argumento semelhante, como um argumento lógico ardiloso pode ser usados para tentar invalidar uma suposta descoberta científica de maneira convincente aos olhos incautos. Os "axiomas" do experimento são os seguintes: 1. Eletricidade e magnetismo são forças. 2. A propriedade essencial da luz é a luminosidade. 3. A Força por seu turno não é nem constitutiva, nem suficiente para a luz. Conclusão: Nem a eletricidade nem o magnetismo são constitutivos ou suficientes para a luz. Se esse argumento fosse lançando por algum crítico de Maxwell antes da descoberta das propriedades paralelas entre a luz a as ondas eletromagnéticas, Maxwell estaria com problemas. O erro aqui reside no segundo "axioma": "A propriedade essencial da luz é a luminosidade". Ora, cientificamente sabemos que esse não é o caso, mas não era isso o que os cientistas sabiam em 1864. Nesse sentido, o argumento se baseia em uma espécie de apelo à ignorância. No caso do argumento de Searle, seria a sintaxe suficiente para a semântica? Filosoficamente, existe uma grande corrente que diz que não, mas isso não é um dado científico, mas uma especulação filosófica e para os Churchlands isso não basta.

No artigo, os Churchlands se dispõe a fazer um breve resumo do conexionismo. Os argumentos deles a respeito do potencial da tecnologia são semelhantes aos que discutimos acima. Ademais, eles também se propõe a discutir a respeito da resposta de Searle ao PDP, chamada de "*Chinese Gym*", onde, em vez de um quarto, teríamos uma academia com dezenas ou centenas de pessoas recebendo paralelamente textos em chinês e os traduzindo simultaneamente. O resultado, segundo Searle, seria tão vazio quanto o do *Chinese Room* convencional. Mas os Churchlands acham essa adaptação ridícula. Primeiro, porque, se levarmos em conta o tecido nervoso humano, estaríamos falando de uma "academia" com mais de oitenta bilhões de pessoas. Segundo, porque não interessa se as pessoas dentro da academia não "entendem" chinês. Já que a intenção é mimetizar o sistema nervoso biológico, leva-se em consideração que um neurônio isoladamente não entende nada daquilo que se passa na mente como um todo, o entendimento se dá em uma esfera sistemática, ou seja, todo o cérebro entende o que se passa, não um neurônio isolado.

É preciso ser justo com Searle. No caso da primeira questão, isto é, a academia com algumas pessoas, a adaptação pode soar absurda considerando que o cérebro possuí bilhões de neurônios, mas o fato é que Searle não estava mirando o cérebro humano quando adaptou experimento, mas sim as redes PDP. Convencionalmente, uma rede PDP tem apenas dezenas ou centenas de unidades, e mesmo hoje, com o advento do *Deep Learning*, não existem nenhuma rede que tenha bilhões de unidades. Em 1990 isso estava ainda mais distante. Portanto, o *Chinese Gym* pode fazer sentido sob esse aspecto.

Já o segundo ponto, isto é, de que a academia entenderia Chinês, é mais ou menos a ressurreição do argumento *System Reply* que Searle considerou em 1980, dessa vez sob a forma de Processamento paralelo distribuído. A resposta de Searle para esse caso será analisada no capítulo seguinte. Por fim, Os Churchlands concluem seu artigo com um apelo à humildade. Não há garantias, segundo eles, de que o conexionismo será capaz de produzir ou sistematizar máquinas com mentes. No entanto, comparado aos modelos simbólicos, se trata de uma evidente evolução. Redes neurais artificiais são capazes de coisas que a IA clássica sequer sonhou e levando em conta a sua idade relativa, tudo o que as Redes Neurais conquistou é apenas o começo. Ademais, eles insistem, o cérebro é um computador, talvez não um computador como os que estamos acostumados, mas certamente é uma máquina de carne que processa informações. O caso para a questão de como esse "computador biológico" concede significado para as coisas que representa ainda é desconhecido,

mas, segundo os Churchlands, não se trata apenas de uma questão linguística ou filosófica, se trata de uma questão que precisa ser descoberta através de um exame neurocientífico, e, tal como na psicologia comum, devemos deixar a investigação meramente linguística de lado e nos concentrarmos na investigação empírica.

5. REDES NEURAIS ARTIFICIAS COMO MODELOS FORMAIS

Searle está ciente a respeito das Redes Neurais artificiais e dos argumentos conexionistas. No artigo de 1990, publicado na mesma edição da *Scientific American* no qual os Churchlands expuseram seus argumentos, Searle, após uma breve explicação do *Chinese Room*, se detém especificamente em lidar com o caso das Redes Neurais Artificiais, em que um programa funciona simulando o cérebro humano, invés de manipular símbolos. Ele também se ocupa em fazer uma refutação do argumento dos Churchlands contra o experimento (No caso, a refutação na qual eles utilizaram o chamado "*Maxwell's Room"*). Essa refutação veremos mais abaixo. A parte em que Searle se ocupa do conexionismo é provavelmente a mais importante do texto e a qual nós exploraremos mais detidamente, porque seus argumentos continuam sendo importantes no debate sobre Inteligência Artificial que é produzida e pesquisada hoje em dia, especialmente no que concernem aos argumentos a favor de representação em redes neurais artificiais.

Para facilitar a compreensão, podemos dividir o argumento em três pontos, bem como expandir a discussão em cada ponto específico. Primeiro, Searle argumenta sobre a natureza dos "Programas de Processamento Paralelo Distribuído", em que ele aponta que qualquer programa que processa suas informações paralelamente pode ser processado linearmente. Nesse ponto, pode ter existido alguma má interpretação de Searle quanto ao termo "Parallel Processing". Originalmente, esse termo foi usado pelos primeiros conexionistas para descrever uma Rede Neural Artificial, sendo cunhado por Rumelhart em seu trabalho de 1986, onde ele descreve, entre outras coisas, que:

> Em nossa opinião, as pessoas são mais inteligentes do que os computadores de hoje porque o cérebro emprega uma arquitetura computacional básica que é mais adequada para lidar com um aspecto central das tarefas naturais de processamento de informações nas quais as pessoas são tão boas. [...] Centenas de vezes por dia nós procuramos por coisas e quase nunca pensamos nesses atos de procura. E, no entanto, cada vez, um grande número de diferentes considerações parece determinar conjuntamente exatamente como alcançaremos o objeto. (RUMELHART, 1986, p. 3)

Os Churchlands tem em mente exatamente aquilo que Rumelhart propunha ao usar o termo "*Parallel Processing*". Hoje, porém, essa conceito caiu em desuso. Searle aparente entendeu "*Parallel Processing"* como uma modificação da arquitetura tradicional de Von Neumann, Tanto que no texto ele usa como exemplo o computador Cray-1. Essa possível má interpretação leva ele ao seguinte argumento: Não importa o cenário, uma Inteligência Artificial simbólica ou sub-simbólica continua apenas replicando a parte formal do nosso comportamento.

Nesse sentido, no segundo ponto da sua argumentação, ele introduz mais uma modificação do *Chinese Room* para lidar com o caso do "processamento paralelo", essa modificação chamamos de "*Chinese Gym"*. No *Chinese Gym,* invés de uma única pessoa trabalhando na manipulação dos símbolos, teríamos dezenas de pessoas que representam as unidades de uma rede neural. Num exercício mais ilustrativo, seria como se, para uma rede neural de duas unidades de entrada, quatro unidades escondidas e duas unidades de saída, teríamos duas pessoas recebendo as palavras em chinês, quatro pessoas fazendo a tradução e selecionando a resposta adequada e outras duas colocando a resposta para fora do quarto. Já dissemos no Capítulo anterior que os Churchlands fizeram troça desse experimento, já que seria um tanto quanto absurdo imaginar uma "academia" análoga aos oitenta e seis bilhões de neurônios do cérebro. Porém, é importante notar, em favor de Searle, que ele provavelmente tinha em mente as primeiras Redes Neurais Artificiais, bem como a sua insistência quanto ao caso de que os experimentos são meramente ilustrativos. Nesse sentido, é possível entender o raciocínio da seguinte forma: mesmo para o caso de redes muito grandes com milhares de unidades, tal como dito antes, a atividade continua sendo um mero funcionamento formal daquilo que nós temos em nossos cérebros.

O terceiro ponto do argumento diz respeito à implementação. Para Searle, nossa mente é um produto do nosso cérebro, isso está derivado diretamente de seu naturalismo biológico. Um programa por sua vez independe do meio, o computador x pode ter uma capacidade de memória maior ou menor do que o computador y, mas supõem-se que o programa deva funcionar normalmente em ambas as plataformas, apesar de uma possível diferença em performance. Não é o caso do cérebro, porque a consciência não é um programa que pode ser retirado de um indivíduo e implementado em outro. As percepções obtidas pelo corpo, bem como os processos bioquímicos influem diretamente no funcionamento dos estados internos do homem. Ainda que esses processos neuronais do cérebro possam ser simulados por um computador, tal como a simulação de um furacão ou de qualquer outro evento natural, a simulação não é o processo neuronal dentro das suas condições naturais (uma vez que o processo neuronal não funciona apenas *per se*), mas apenas uma reprodução das suas condições substanciais. O cérebro não simula ou formaliza uma atividade neuronal, ele efetivamente tem atividade neuronal que, segundo a argumentação do próprio Searle, causa os eventos mentais internos.

Nesse ponto, chegamos ao aspecto mais importante. Se os neurônios têm, comprovadamente, uma função ativa, se nosso comportamento é, comprovadamente, gerado por essa atividade, porque um modelo matemático de um neurônio, como o Perceptron de Rosenblatt, não pode também gerar

pensamento? Justamente porque a formalização da ação "*all-or-nothing"* teorizada por McCulloch e Pitts não é mais do que uma formalização substancial de parte da atividade de um neurônio. Um neurônio artificial tem entradas, saídas, função de ativação, pesos, somatório, todas essas características são inspiradas na formulação do neurônio biológico que existia nos anos 1940s, e, por mais fidedignas que sejam, sua razão de existir não é serem idênticos aos neurônios biológicos, mas replicarem sua função comportamental. Nesse sentido, alguns pesquisadores de IA possuem descrições dos neurônios artificiais que favorecem a Searle. Como já descrito, até mesmo o refinamento daquilo que faz um neurônio artificial funcionar de maneira mais aproximada ao neurônio do cérebro humano, isto é, sua função de ativação, não é mais do que um modelo matemático, alcançado através de testes de eficiência em certas atividades.

Alguém pode, por sua vez, perguntar: "Mas o que aconteceria se criássemos uma rede neural artificial com o mesmo número de neurônios que o cérebro e com a mesma distribuição na arquitetura de funcionamento?", em resumo, o que aconteceria se fizéssemos uma rede com 86 bilhões de neurônios? Segundo o argumento de Searle, essa rede funcionaria como um zumbi. Suas atividades até poderiam ser idênticas a de um ser humano de carne e osso, mas o seu "funcionamento interno" careceria de semântica. Esse problema é conhecido como "*Philosophical Zombie*".

Antes de explorarmos o que Searle tem a dizer sobre o *Philosophical Zombie*, é importante vermos antes a resposta de Searle ao argumento *Maxwell's Room* dos Churchlands.

Em miúdos, no capítulo anterior vimos que o experimento do "*Maxwell's Room*" visa contrapor o *Chinese Room* a partir de uma possível sugestão de que Searle falha em considerar uma conexão causal entre sintaxe e semântica que supostamente ainda não foi descoberta cientificamente. Searle contrapõe essa linha de pensamento com o seguinte raciocínio, a analogia com as descobertas de Maxwell não se sustentam porque a luz e o electromagnetismo são dois fenômenos naturais observáveis, naturalmente, por esses motivos suas relações causais foram posteriormente descobertas. A sintaxe, ao contrário, não é um fenômeno natural e um objetos abstratos, por esse motivo, não pode ter alguma propriedade que pode ser investigada empiricamente.

> A analogia com os símbolos formais falha porque os símbolos formais não têm poderes causais físicos. O único poder que os símbolos têm, enquanto símbolos, é o poder de causar a próxima etapa do programa quando a máquina está funcionando. (SEARLE, 1990, p. 30)

5.1 ZUMBI FILOSÓFICO

O problema do zumbi não nasceu na discussão de Inteligência Artificial, e não necessariamente se trata de um argumento para lidar com ela. Conceitualmente, um Zumbi filosófico é um ser humano aparentemente normal, que se comporta como qualquer outro, mas que por ser um zumbi carece de atividade introspectiva. Nesse sentido, não há um "*What Is It Like"* quando se fala de um zumbi. O problema foi levantado por Keith Campbell em seu livro *Body and Mind* de 1970, sob nome de "*Imitation Man*". Em 1974, Robert Kirk reintroduziu o problema, dessa vez sob nome de "Zombie" e mais tarde, David Chalmers deu prosseguimento a discussão, atualizando seus conceitos. Parte da discussão que se faz hoje do Zumbi filosófico vem do trabalho de Chalmers. Essencialmente, o Zumbi filosófico foi criado para lidar com o fisicalismo, uma vez que o experimento prevê uma figura de carne e osso como um ser humano, e não apenas um cérebro ou sistema nervoso. Mas é possível fazer uma adaptação do problema para o caso das redes neurais artificiais, uma vez que elas pretendem ser um modelo artificial do sistema nervoso biológico. Searle nunca tratou do problema do Zumbi original de maneira relevante, no entanto, em seu livro *The Rediscovery of Mind* de 1992, há uma série de experiências que visam lidar justamente com o caso de um "Robô humanoide com neurônios artificiais".

Grande parte dos argumentos que ele levanta nesse livro são muito semelhantes aos levantados no artigo de 1990, mas as diferença são onde Searle leva esses argumentos. Dissemos que alguém poderia sugerir que uma rede neural com 86 bilhões de neurônios artificiais poderia produzir uma consciência como a nossa. Em certa parte do livro, Searle sugere um experimento em que, em resumo, médicos substituem os neurônios biológicos de um paciente por neurônios artificias perfeitamente fabricados. O experimento funciona assim: A medida que vão substituindo os neurônios artificiais, os médicos perguntam ao paciente se ele continua enxergando, sentindo, percebendo o mundo ao seu redor. Externamente, as respostas são sempre afirmativas, mas internamente, o paciente tenta, com todas as forças, dizer que está perdendo os sentidos, porém seu comportamento não reage da maneira como ele deseja. Quando todos os neurônios são substituídos, o paciente se comporta como um ser humano normal, mas está mentalmente morto. Essa situação está de acordo com o que foi argumentado anteriormente. Para Searle, não importa a quantidade de unidades de uma rede neural, tampouco o quão eficaz é a sua função de ativação e sua capacidade em identificar e repetir padrões de comportamento, uma vez que um neurônio artificial só é capaz de replicar a atividade formal de uma neurônio biológico, não há qualquer atividade interna.

Alguém ainda pode argumentar: “Se o comportamento é idêntico ao de um ser humano, logo, deve haver consciência, não?” A questão é que, conforme argumento em seu livro de 1992, Searle considera que levar em conta apenas o comportamento no que diz respeito as experiências mentais internas é um exercício incompleto, senão errôneo. Entre os argumentos que ele descreve para fundamentar essa afirmação está o caso da Síndrome de Guillain-Barré. Essa condição se trata de uma franqueza muscular causada por um ataque do sistema imunitário ao sistema nervoso periférico. Em certos casos o corpo fica inteiramente paralisado, mas atividade consciente permanece operante. A paralisia decorre da inoperação do sistema nervoso motor. Além disso, existe um ponto importante quando consideramos a experiência interna em outros indivíduos além de nós mesmos. Apesar de o comportamento ser um fator importante para considerarmos que existe consciência além da nossa própria consciência, a própria fisiologia das outras pessoas e dos animais que nós cercam nós levam a concluir, segundo Searle, que existe atividade cerebral além da nossa. Uma vez que o outro (e não precisa ser necessariamente um humano) tem olhos, boca, pele, respira e etc, essa proximidade torna a conclusão natural. Por esse motivo, o Problema das outras mentes não passa de um problema hipotético. Na prática, qualquer ser humano, e até mesmo certos animais, considera que existe consciência em outros indivíduos.

Como vimos no primeiro capítulo, Turing considerava que o comportamento era suficiente para determinar a existência ou não de pensamento. Um dos seus argumentos estava embasado na questão behaviorista. Conforme argumentou em “*Computer Machinery and Intelligence”*, Turing dizia que a comunicação era única forma de concluirmos que outra coisa além de nós era capaz de pensar. Em seu tempo, ele não levou em consideração argumentos semelhantes aos de Searle quanto à questão fisiológica e causal do funcionamento do cérebro. Em sua defesa, argumentos como esse ainda não existiam na psicologia feita na época, e o behaviorismo era a teoria do momento nos institutos de psicologia.

6. CONSIDERAÇÕES FINAIS

A discussão não avançou muito em aspectos teóricos após 1994. Searle ainda viria a tratar do conceito em trabalhos posteriores, mas o núcleo do seu pensamento foi posto nos artigos de 1980, 1990 e nos livros de 1992 e 1997. Do lado conexionista, apesar dos livros dedicados a expandir a discussão para a audiência em geral, escritos por Patricia e Paul Churchland em conjunto com os fundadores da área como Sejnowski e McCelland, nenhum dos dois adicionou pormenores teóricos extras à discussão. Provavelmente o texto mais interessante escrito após os anos 1990s, foi um artigo de Searle para o *The Wall Street Journal* intitulado "*Watson Doesn't Know It Won on 'Jeopardy!"*, em que ele argumenta, usando as mesmas ferramentas que havia usado em seus artigos sobre Inteligência Artificial Forte, que o supercomputador IBM Watson não é capaz de pensar.

O tempo, entretanto, esteve em favor da tecnologia. Ainda que hoje continue uma tarefa dura defender uma Inteligência Artificial pensante, o conexionismo viu o surgimento de animadores avanços com o *Machining Learning* e o *Deep Learning*, além de um interesse massivo das grandes corporações nessa forma de computação. Companhias como Amazon e Google gastam dezenas de milhões de dólares anualmente para refinarem seus algoritmos de IA e implementarem novas tecnologias de reconhecimento de padrões de comportamento dos usuários dos seus serviços. A palavra da moda dentro da Ciência da Computação ultimamente tem sido "*Deep Learning"*. Com o *Deep Learning*, os conexionistas de trinta anos atrás estão podendo ver recentemente o surgimento de programas capazes de realizar tarefas que eles advogavam ainda nos anos 1980. Diferente de uma rede neural "comum", (hoje em dia apelidada de "*Shallow Neural Network*"), uma rede *Deep Learning*, como seu nome sugere, possuí dezenas a centenas de camadas de unidades escondidas que fazem um "aprendizado profundo" dos seus inputs. O resultado é extremamente fino. Se em 1990 uma rede neural do tipo ADALINE era capaz de distinguir, em uma fotografia, o que era um gato e o que era um cachorro, hoje, uma rede neural de aprendizado profundo é capaz de distinguir vários objetos em uma cena, bem como ser capaz de classificar objetos como uma precisão equivalente ou superior ao dos seres humanos, além de ser capaz também de gerar textos, reconstruir e gerar imagens e processar linguagem natural. Essas redes estão tão introduzidas dentro da industria que é particularmente difícil pensar no futuro sem elas. Novas questões surgem, como as questões éticas, bem como, devido ao poder dessas redes, ressurge a pergunta: "Pode uma máquina pensar?"

Atualmente há pouca filosofia que lide com as redes do tipo *Deep Learning*. Provavelmente, isso se deve ao fato de ser um campo muito recente, cujo os impactos ainda estão sendo avaliados, mas

aparentemente isso está prestes a mudar. Algumas Inteligências Artificiais já se propõem a serem inteligências "gerais". Elas não são "agentes", uma vez que não tem um objetivo específico, mas se adaptam para realizar um certo número de tarefas. Entre as mais notáveis está o programa GPT-3. Lançado em 2015, o programa tem a proposta de produzir textos como um ser humano. Seu desempenho em conversas naturais é algo nunca antes visto, e o programa é capaz inclusive de gerar textos que se assemelham a poesia. Alguns filósofos fizeram alguns comentários sobre o programa, as visões foram mistas. David Chalmers elogiou o GPT-3, embora tenha destacado que, ao contrário do que pode parecer a princípio, ele "não pensa".

> O GPT-3 é instantaneamente um dos sistemas de IA mais interessantes e importantes já produzidos. […] Parece estar mais perto de passar no teste de Turing do que qualquer outro sistema até hoje […] mas o treinamento do GPT-3 é irracional. Se trata apenas de analisar estatísticas da linguagem. (CHALMERS, 2020, Disponível em: < https://dailynous.com/2020/07/30/philosophers-gpt-3/ >)

Luciano Floridi, por sua vez, fez alguns comentários mais criticos em relação ao programa. Filósofo preocupado com as questões éticas da informação, ele diz:

> Existem algumas consequências significativas quanto a industrialização da produção automática e barata de bons artefatos semânticos. (FLORIDI & CHIRIATTI, 2020, p. 681)

Por mais poderosos que sejam, os recentes programas criados com *Deep Learning* ainda não tem o suficiente para convencer a maioria dos cientistas e filósofos de que possuem alguma atividade interna, ainda que existam aqueles que argumentem o contrário. O artigo de Searle para a *Scientific American* já tem ao menos trinta e dois anos, esses trinta e dois anos, como esperado, estiveram a favor do avanço técnico da Inteligência Artificial, porém, seja um algoritmo de Inteligência Artificial Simbólica simples, seja uma rede neural com milhares de unidades, a Inteligência Artificial continua sendo boa apenas em copiar o nosso comportamento, mas não o que há por trás dele.

REFERÊNCIAS

ABBAGNANO, Nicola, Dicionário de Filosofia, Em: Fenomenologia, Martins Fontes, 1998

BUCKNER, Cameron, GARSON, James, Connectionism, The Stanford Encyclopedia of Philosophy (2019), URL = <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>, Acesso em Outubro de 2022.

CARVALHO, Joelma Marques, Searle e os Desafios da Inteligência Artificial Forte, Revista Reflexões, Ano 10, Nº 18, 2021

CHALMERS, David, "GPT-3 and General Intelligence". Daily Nous. Philosophers On GPT-3 (updated with replies by GPT-3). (2020) URL=<https://dailynous.com/2020/07/30/philosophers-gpt-3/#chalmers> Acesso em Dezembro de 2022.

CHURCHLAND, Paul M., CHURCHLAND, Patricia S., Could a machine think?, Scientific American, 262, 1990.

CHURCHLAND, Paul M., Some Reductive Strategies in Cognitive Neurobiology, Mind, Volume XCV, Issue 379, Pages 279–309, 1986

DREYFUS, Hubert, Alchemy and AI, RAND Corporation, 1965;

DREYFUS, Hubert, What Computers Can't Do, New York: MIT Press, 1972;

FLORIDI, Luciano, CHIRIATTI, Massimo. "GPT-3: Its Nature, Scope, Limits, and Consequences". Minds and Machines. 30 (4): 681–694, 2020

HARNAD, Stevan, The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence, 2004

HAYKIN, Simon, Redes Neurais: Princípios e Práticas, Bookman Editora, 2007;

HONDERICH, Ted (ed.). The Oxford Companion to Philosophy New York: Oxford University Press, *In*: Phenomenology, 1995

HORGAN, Terence; TIESON, John. Connectionism and the Philosophy of Mind, Springer Netherlands, 1991;

JACOB, Pierre, Intentionality, The Stanford Encyclopedia of Philosophy (2019), URL = <https://plato.stanford.edu/archives/win2019/entries/intentionality/>, Acesso em Outubro de 2022.

JEFFERSON, G. The Mind of Mechanical Man. British Medical Journal, 2(4568): 211–213, 1948

KEELEY, Brian L., Paul Churchland, Cambridge University Press, 2006

KIRK, Robert, Sentience and Behaviour, Mind, vol. 83, pp. 43–60, 1974

KIRK, Robert, Zombies, The Stanford Encyclopedia of Philosophy (2021), URL = <https://plato.stanford.edu/archives/spr2021/entries/zombies/>, Acesso em Dezembro de 2022.

McCAULEY N. Robert, The Churchlands and Their Critics, Blackwell Publishers, 1994

McCORDUCK, Pamela, Machines Who Think (2nd ed.), Natick, MA: A. K. Peters, Ltd, 2004

MERLEAU-PONTY, Maurice, Phenomenology of Perception, new trans. by Donald A. Landes, Routledge, 2012

RESCOLA, Michael, The Computational Theory of Mind, The Stanford Encyclopedia of Philosophy (2020), URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>, Acesso em Outubro de 2022.

RUMELHART, David E; McCLELLAND, JAMES L. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, 1987.

RUSSELL, Stuart; NORVIG, Peter. Artificial Intelligence: A Modern Approach (2nd ed.), 2003

SAYGIN, A. P.; CICEKLI, I.; AKMAN, V., Turing Test: 50 Years Later, Minds and Machines, 10 (4): 463–518, 2000

SEARLE, John, Minds, Brains and Programs, Behavioral and Brain Sciences, 3 (3): 417–457, 1980

SEARLE, John R, Intentionality, an essay in the philosophy of mind, Cambridge [Cambridgeshire]: Cambridge University Press, 1983

SEARLE, John, Is the Brain's Mind a Computer Program?, Scientific American, vol. 262, January 1990

SEARLE, John, Is the Brain's Mind a Computer Program?, Scientific American, vol. 262, 1990a

SEARLE, John. "Watson Doesn't Know It Won on 'Jeopardy!'". The Wall Street Journal. (2011). URL = <https://www.wsj.com/articles/SB10001424052748703407304576154313126987674>, Acesso em Dezembro de 2022.

SEARLE, John, The Rediscovery of the Mind, A Bradford Book, 1992

SMITH, David W., Phenomenology, The Stanford Encyclopedia of Philosophy (2018), URL =<https://plato.stanford.edu/archives/sum2018/entries/phenomenology/>, Acesso em Outubro de 2022.

SMITH, Barry, John Searle, Cambridge University Press, 2003

SILVA, Ivan Nunes, SPATTI, Danilo Hernane, FLAUZINO, Rogério Andrade, LIBONI, Luisa Helena Bartocci, Alves, Silas Franco dos Reis, Artificial Neural Networks: A Practical Course, Springer International Publishing, 2016

TURING, Alan, Computing Machinery and Intelligence, Mind, LIX (236): 433–460, Outubro 1950

TURING, Alan, Computing Machinery and Intelligence. In: EPSTEIN, R., ROBERTS, G., BEBER, G. (eds) Parsing the Turing Test. Springer, Dordrecht, 2009