Revista Eletrônica de Filosofia
Philosophy Eletronic Journal
ISSN 1809-8428

São Paulo: Centro de Estudos de Pragmatismo
Programa de Estudos Pós-Graduados em Filosofia
Pontifícia Universidade Católica de São Paulo
Disponível em http://www.pucsp.br/pragmatismo

Vol. 13, nº. 2, julho-dezembro, 2016, p.152-159

MÁQUINAS PODEM SE TORNAR CONSCIENTES?

Nara Ebres Bachinski
Mestranda do Programa de pós-graduação da Universidade Federal de Santa Maria
naraebresb@gmail.com

Resumo: Este artigo tem como objetivo apresentar alguns pontos centrais da discussão acerca da possibilidade de consciência de máquina. Apresento primeiramente o conceito de inteligência artificial introduzido por Alan Turing na década de 1950 e, depois, a objeção de John Searle aos programas de inteligência artificial forte, bem como as réplicas de Jerry Fodor, de Paul e Patricia Churchland a John Searle e, por fim, as perspectivas para consciência de máquina a partir das redes neurais artificiais.

Palavras-Chave: Inteligência artificial. Redes neurais artificiais. Filosofia da mente. Consciência de máquina.

CAN MACHINES BECOME CONSCIOUS?

Abstract: *This paper aims to present some central points of the discussion about the possibility of machine consciousness. First, I present the concept of artificial intelligence introduced by Alan Turing in the 1950s, and then John Searle's objection to the programs of strong artificial intelligence, including the replicas of Jerry Fodor, Paul and Patricia Churchland to John Searle and, finally, the perspectives for machine consciousness to artificial neural networks.*

Keywords: *Artificial intelligence. Artificial neural networks. Philosophy of mind. Machine consciousness.*

* * *

Introdução

Em meados do século XX, com o desenvolvimento dos primeiros computadores modernos, alguns cientistas e filósofos começaram a vislumbrar a possibilidade de efetivamente se construir máquinas pensantes. Alan Turing tratou do problema em meados dos anos 1950, mas considerou inadequado o uso da palavra "pensamento" nesse contexto.[1] Essa palavra é ambígua nos seus usos comuns, e não tem propriamente usos técnicos nas ciências. Pode ser usada de

[1] Alan Turing, Maquinário computacional e inteligência (1950)

várias maneiras, para descrever fenômenos muito distintos, tais como calcular, planejar, lembrar, imaginar etc. Para contornar essas dificuldades, Turing propôs que o problema fosse tratado de modo comportamental.

Alguns autores mais recentes (por exemplo, Massimo Pigliucci[2]) tratam a palavra "pensamento" nessa questão como um termo genérico que engloba ao menos dois subconjuntos: inteligência e consciência. Estes termos também sofrem das ambiguidades da palavra "pensamento", mas são mais específicos. Em um dos vários sentidos da palavra "inteligência" – o de calcular –, computadores digitais são obviamente inteligentes. Mas quando se trata de pensar no sentido de ser consciente, não é nada claro se algum dia eles terão essa capacidade. Com relação a essa questão mais específica, a de se computadores podem ser conscientes ou vir a ser, há uma divisão bem aguda na literatura. Alguns parecem seguir a perspectiva de Turing e afirmam que na melhor das hipóteses podemos ter critérios comportamentais da inteligência e da consciência.[3] Outros dizem que critérios comportamentais são suficientes para a atribuição de inteligência e consciência.[4] E ainda há os que defendem que algo mais é exigido, ao menos para a atribuição de consciência, como indicadores biológicos da consciência ou alguma explicação de tipo físico (a mera exibição de comportamentos compatíveis seria insuficiente).[5]

Nos últimos anos, uma das linhas de pesquisa mais promissoras nesta área tem sido a das chamadas "redes neurais artificiais". Trata-se de pesquisas em inteligência artificial que projetam conceitualmente e redigem programas de computador que funcionam em máquinas com arquiteturas que procuram imitar a estrutura de um cérebro humano. Diferente dos programas de inteligência artificial clássicos, essas estruturas têm capacidades de aprendizagem e estruturas de processamento complexas (hierárquicas, paralelas e bidirecionais). Redes neurais artificiais podem, assim, apresentar modelos de aprendizagem semelhantes ao de redes neurais biológicas.[6] Assim como as redes neurais biológicas, as redes artificiais também possuem nódulos que são análogos a "neurônios", com conexões de entrada e de saída que podem ser ajustadas de acordo com os *inputs* e capazes de aprendizagem (ver Fernandes, 2005).

Este artigo apresenta quatro pontos centrais dessa discussão: (i) o conceito de inteligência artificial introduzido por Turing na década de 1950, (ii) a objeção de Searle aos programas de inteligência artificial forte, (iii) as réplicas de Fodor e dos Churchlands a Searle e (iv) as perspectivas de inteligência artificial em redes neurais artificiais.

1. **O teste de Turing**

A discussão acerca da Inteligência Artificial está centrada nas questões "pode uma máquina pensar?"e "pode uma máquina imitar o pensamento humano?". Por

[2] Ver, por exemplo, a entrevista disponível em: <https://youtu.be/t7vG2WYuWks>.

[3] John Searle, A mente do cérebro é um programa de computador? (2010)

[4] Keith Frankish & William Ramsey, The Cambridge handbook of artificial intelligence (2014).Ver, especialmente, capítulo 3.

[5] Patrícia S. Churchland, Poderia uma máquina pensar? (2015)

[6] Para uma caracterização mais extensa das redes neurais artificiais, ver Fernandes (2005, p. 59).

volta de 1950, Alan Turing levanta essas questões de um modo indireto. A palavra "pensamento" parece-lhe ser demasiadamente vaga para permitir uma resposta clara. Contudo, pode-se tratar de questões parecidas ou, ainda, que possam lançar alguma luz sobre esse tema. Como um recurso para tentar solucionar essas questões, Turing propõe um teste, o qual ele denomina "jogo da imitação". Nesse jogo, haveria três indivíduos, um homem, uma mulher e um interrogador, cada qual em uma sala isolada, sem comunicação direta. O interrogador tentaria descobrir em qual sala está o homem e em qual sala está à mulher por meio da formulação de perguntas formuladas escritas e/ou datilografadas. O objetivo do homem seria tentar induzir o interrogador ao erro, e o da mulher o de tentar levá-lo ao acerto. Vence o jogo quem atingir seu objetivo. Diante disso, Turing levanta a questão: o que aconteceria se o homem fosse substituído por um computador? Este seria capaz de enganar o interrogador com tanto sucesso quanto o homem originalmente conseguira? Essas questões passariam a substituir a questão original sobre se máquinas poderiam pensar. Se fosse possível a um programa de computador ser tão bem sucedido nesse jogo quanto um homem, então essa máquina passaria no (hoje chamado) "teste de Turing". Esse computador seria capaz de imitar o comportamento humano de modo tão eficaz que poderia ser frequentemente confundido com um ser humano.

Esse teste é puramente comportamental, e não se propõe a responder diretamente à pergunta sobre se máquinas podem pensar ou ser conscientes. Embora o próprio Turing tenha indicado que essa questão não tem como ser respondida, dada a vagueza das palavras com as quais é formulada, diversos outros filósofos se propuseram a tratar dela diretamente, defendendo respostas positivas ou negativas, conforme veremos.

Turing sugere que a pesquisa com máquinas que imitam o comportamento "pensante" de seres humanos poderia usar como modelo as etapas do desenvolvimento da mente de uma criança, considerando: "(a) estado inicial da mente ao nascer; (b) a educação que recebeu; (c) outras experiências, que não são descritas como educação, a que foi submetida" (Turing, 1950). Entretanto, no desenvolvimento de uma máquina se excluiria o método *afetivo* de aprendizagem. O aprendizado de máquina seria baseado em "[...] um sistema completo de inferência lógica embutida. Nesse caso, a memória seria ocupada em grande parte por definições e proposições. As proposições seriam de vários tipos; por exemplo, fatos bem estabelecidos, conjeturas, teoremas matemáticos demonstrados, enunciados de autoridade, expressões que tenham a forma lógica de proposição, mas não de valor-crença" (Turing, 1950). Desse modo, o desenvolvimento da máquina assemelhar-se-ia ao desenvolvimento humano e poderia, ao final, resultar na capacidade de imitar o comportamento humano. Esse modelo de IA com aprendizado veio a ser ponderado e valorizado por cientistas de computação (ver Starzyk e Prasad, 2011) no desenvolvimento em modelos de redes neurais artificiais.

2. **O quarto chinês**

Diversos filósofos objetaram que a expectativa gerada pelos programas de inteligência artificial não teriam como em princípio ser satisfeita. Com certeza, computadores são capazes de calcular. Entretanto, aquilo que chamamos de pensamento e consciência (cujos modelos são os humanos) exigiria algo mais. Um

dos filósofos que se destacou por argumentar nesse sentido foi John Searle. O artigo “A mente do cérebro é um programa de computador?”diferencia entre inteligência artificial forte e fraca. Os programas de pesquisa em IA fraca compreendem a IA como instrumentos para o estudo da mente, uma vez que permitem a formulação e teste de hipóteses de uma maneira útil e precisa. Já os programas de pesquisa em IA forte concebem a IA não são apenas como instrumentos de pesquisa, mas como modelos da própria mente. No artigo mencionado, Searle criticou a concepção de IA forte por meio do experimento mental do “quarto chinês”. Neste experimento há uma pessoa em um quarto onde há duas aberturas, uma de entrada, outra de saída. Nesse quarto, essa pessoa recebe, pela abertura de entrada, histórias escritas em chinês. Dentro do quarto há um manual também em chinês, mas com as regras em inglês (língua nativa dessa pessoa). Este manual diz coisas do tipo: se você receber uma folha escrita com os sinais tais e tais, vá para a página X deste manual e copie o que está lá escrito em uma nova folha e a entregue pela abertura de saída. Trata-se, naturalmente, de um manual muito grande, que prevê um número alto de possibilidades combinatórias de sinais. Mas, efetivamente, pela abertura de entrada são inseridas frases chinesas e pelo de saída, frases também em chinês que respondem ou contestam ou comentam ou reagem ao que foi inserido, na forma de um diálogo em chinês. A pessoa que está dentro do quarto, no entanto, nada sabe de chinês, e apenas manipula símbolos sem saber o que significam. Segundo Searle, essa pessoa faz o mesmo que um programa de computador, que para cada *input* determinado e fornece um *output*. Ora, do mesmo modo que a pessoa dentro do quarto não sabe chinês, do mesmo modo também o programa de computador não sabe o significado dos sinais que recebe como input ou que emite como output. Para Searle, a instanciação de um programa de computador, por consistir na mera operação formal com símbolos, não é suficiente para produzir estados mentais.

Não precisamos nos limitar a experimentos mentais para apresentar o ponto de Searle. No início do século XXI, o programa ELIZA foi projetado para produzir respostas simples e minimamente inteligíveis, coerentes e corretas a perguntas que recebia. Esse programa era capaz de manter um breve diálogo com um ser humano adulto que falasse a mesma língua que ELIZA. Segundo Searle, por mais parecido que programas desse tipo possam ser com o comportamento humano, não há neles propriamente pensamento ou compreensão, pois quando um ser humano pensa ou compreende algo, não está apenas processando ou manipulando sinais, mas é capaz de ligá-los cognitivamente aos objetos no mundo. O sinal representa algo no mundo para quem o compreende, ou desempenha alguma outra função semântica. Essas relações semânticas estão ausentes nos programas de computador. Para Searle, a mera instanciação de um programa de computador, portanto, não produz estados mentais. Computadores possuem apenas sintaxe e não semântica.

O argumento de Searle foi muito influente e é seguidamente visto como uma das objeções mais fortes aos programas de inteligência artificial forte. Críticas igualmente importantes ao argumento de Searle podem ser encontradas em Fodor e Patricia Churchland.

3. **Jerry Fodor**

Em resposta a Searle, Fodor(2010) concede que de fato a instanciação de um programa não é suficiente para que haja estados mentais (a mera manipulação de

símbolos não constitui um estado mental). Para que haja estados mentais é necessário que símbolos façam referência a algo no mundo (isto é, que tenham intencionalidade e semântica). Contudo, segundo Fodor, o experimento mental de Searle mostraria apenas que a conexão causal entre os símbolos e as coisas que Searle imagina haver não é do tipo certo. Ambas as questões estão ligadas, uma vez que para admitir um estado mental intencional é necessário que estados mentais tenham uma relação adequada (isto é, intencional) com objetos referidos pelos sinais.

Fodor argumenta que se pensarmos no experimento do quarto chinês como um software e a ele ligarmos um robô com características perceptuais, então o robô teria as relações causais do tipo adequado com os objetos referidos pelos sinais, pois a manipulação de símbolos estaria causalmente conectada a objetos no mundo.

Para Fodor, é razoável supor que o tipo certo de relação causal é o que ocorre entre o cérebro e os objetos da percepção, ou entre cérebros e objetos distantes. Todavia, disso não se segue que apenas os cérebros possam estar nessas relações e que possuir um cérebro biologicamente parecido com o nosso seja condição necessária para que haja esse tipo de relação. Tampouco se seguiria que manipulações formais de símbolos estejam entre essas relações. Dessa maneira, se algum robô possuir um software acoplado a um aparato sensório e ele associar adequadamente símbolos com objetos no mundo, então esses símbolos seriam significativos semanticamente do mesmo modo que as palavras que pensamos com nossos cérebros são para nós.

Searle rebate essa objeção dizendo que um robô com o aparato sensório e um software adequado ainda não é suficiente para ter as relações causais do tipo certo. Para ele, é necessário que o robô tenha *consciência* da relação causal entre o símbolo e o objeto e que essa relação possa produzir algum conteúdo intencional que possibilite o surgimento de uma memória, ou uma crença, ou uma experiência visual ou ainda uma interpretação semântica de alguma palavra.

4. **Os Churchlands**

Outra linha de argumentação contrária a Searle foi apresentada por Paul e Patricia Churchland em “Uma máquina poderia pensa?”(2015). Eles concordam com Searle e Fodor que a mera manipulação simbólica não pode constituir estados mentais. Contudo, defendem que manipulações simbólicas com um grau de complexidade adequado e conexões adequadas ao meio podem produzir estados mentais em alguns casos. Nesse sentido, sugerem o desenvolvimento de máquinas de inteligência artificial com arquiteturas que imitam o cérebro humano. Segundo eles, o argumento de Searle não demonstraria o que pretende, uma vez que uma de suas premissas seria injustificada. Eles reconstroem o argumento de Searle assim:

Premissa 1: programas de computadores são formais (sintaxe).

Premissa 2: mentes humanas têm conteúdo mental (semântica).

Premissa 3: a sintaxe não é constitutiva nem suficiente para a semântica.

Conclusão: programas de computador não são nem constitutivos nem suficientes para mentes.

O problema está na Premissa 3, que torna o argumento circular (petição de princípio). Desse modo, o argumento não mostra que não é possível que *softwares* mais complexos não possam produzir produtos semânticos. Uma possível solução para esse problema seria programas de pesquisa em redes neurais artificiais.

Diante disso, os Churchlands propõem um modelo de rede neural para inteligência consciente. Esse modelo consiste em uma rede com três camadas, cada qual com unidades conectadas em paralelo com unidades da próxima camada. Quando uma camada for ativada por um *input* ela produzirá um estímulo que transmitirá para suas conexões e consequentemente para todas as outras camadas. O mesmo ocorrerá com *outputs*. Entretanto, no estímulo de *input* haverá a ocorrência de inúmeros vetores de entrada, esses vetores serão convertidos em um só vetor de saída. Esse modelo, segundo eles, não é ameaçado pelo argumento de Searle, uma vez que sistemas em paralelo não manipulam símbolos da mesma forma que computadores digitais o fazem.

5. **Discussão e hipóteses**

A discussão anteriormente apresentada sugere que a manipulação simbólica não era suficiente para a emergência de estados mentais (Searle). Era necessário algo mais para que isso ocorresse, como aparatos sensórios (Fodor) e máquinas que imitassem a arquitetura do cérebro humano (Churchlands). Strarzik e Prasad (2011) propuseram um modelo que contemplasse essas exigências e idealizaram uma rede neural artificial com sensores que possivelmente identificam coisas no mundo. Essa máquina teria inicialmente a capacidade de aprendizado similar à de uma criança – como Turing (1950) sugeriu. Não obstante, para que isso aconteça, é necessário uma inteligência corporificada, ou seja, um corpo mecânico equipado a sensores capazes de detectar e reconhecer objetos, assim como aprender os efeitos de suas ações. Dessa maneira, a máquina estará ciente de si mesma, da mesma maneira que estará ciente de outros objetos.

O modelo proposto por Strarzyk e Prasad consiste em uma arquitetura modular composta por três blocos funcionais: sensório-motor, executivo central, e memória episódica e aprendizagem. Cada bloco é composto por outras partes com funções específicas.

A máquina necessita, para que seu aprendizado seja eficiente, de um mecanismo de *atenção* e de *alteração de atenção*. Esta última corresponde a um processo dinâmico resultante da concorrência entre as representações relacionadas com motivações, *inputs* sensoriais e pensamentos internos, incluindo sinais espúrios. Assim, a alternação da atenção resulta de uma experiência cognitiva deliberada ou pode resultar de processo subconsciente (estimulada por sinais internos ou externos). Dessa maneira, embora a *atenção* seja uma experiência consciente, a *alteração da atenção* não tem de ser (2010, p. 8). Esses autores acreditam que “uma máquina só se tornará consciente uma vez que possuir mecanismos necessários para a percepção, ação, aprendizagem, memória associativa e possuir um executivo central que controla todos os processos(consciente ou inconsciente) da máquina; o executivo central é guiado pela motivação da máquina, seleção de metas, alteração de atenção, memória semântica e episódica e usa a percepção cognitiva e compreensão cognitiva das motivações, pensamentos ou planos para controlar a aprendizagem, atenção,

motivações, e acompanhamento das ações" (2010, p. 9)[7]. É a partir do executivo central relacionando experiências cognitivas de motivações e planos que pode surgir à autoconsciência em uma máquina.

Contudo, não há consenso entre os estudiosos de IA acerca do desenvolvimento de consciência através de redes neurais artificiais. Esta é uma questão que continua em aberto.

6. **Redes neurais artificiais**

Os primeiros modelos de redes neurais artificiais foram propostos na década de 1940.[8] Em 1957, Frank Rosenblatt criou a primeira rede neural artificial com o nome de *Perceptron*, que foi posteriormente abandonado devido a sua incapacidade de processar disjunções exclusivas e decidir entre duas ou mais variáveis.[9]A partir da década de 1980, com o aumento substancial e rápido da capacidade de processamento dos computadores, as pesquisas em redes neurais artificiais avançaram na simulação de redes neurais. O desenvolvimento posterior da neurociência, da engenharia de computação e da robótica permitiu a elaboração de modelos de redes neurais artificiais baseados na arquitetura do cérebro humano, com redes neurais artificiais modulares.[10]Uma rede neural modular contém módulos distribuídos em camadas, cada um dos quais com funções específicas e processamento autônomo relativamente a outros módulos (memória de longo prazo, memória episódica, atenção, cálculo, decisão, planejamento etc.), conectados hierarquicamente. Nos últimos anos, houve grande interesse de estudiosos pelo desenvolvimento desse tipo de modelo. Um exemplo disso encontra-se em Strarzyk&Prasad (2011), que apresentam um modelo hierárquico de processos no qual a informação flui através de módulos funcionais.[11]Diferente da rede neural artificial *perceptron*, a rede neural artificial baseada em módulos não opera somente em fluxos hierárquicos, mas também em fluxos bidirecionais, além de possuir mais de uma camada intermediária. Dessa maneira, pode processar maior quantidade de *inputs* e possibilita *outputs* mais específicos.

7. **Considerações Finais**

O presente artigo pretendeu mostrar de forma breve os argumentos acerca da IA e da emergência de consciência através de redes neurais artificiais. Esses argumentos ainda estão em discussão e os modelos de redes neurais artificiais estão em pleno desenvolvimento, tanto no aspecto técnico quanto no aspecto conceitual. Novas tecnologias nessa área demandam reflexão e possível elaboração de novas hipóteses e teorias mais globais que tornem possível o avanço seguro dos projetos de pesquisa. Um dos objetos principais dessas reflexões e hipóteses são as aplicações das redes neurais artificiais em novas tecnologias (na computação e na

[7] Traduzido por mim.

[8] Ver McCulloch&Pitts (1943).

[9] Ver Rosenblatt (1958) e Minsky&Papert (1969).

[10] Ver Rumelhart&McClelland (1986).

[11] Módulo funcional é uma parte que compõe o *software* no qual se encontram as repartições de memória, atenção, memória episódica.

robótica) e o estudo de modelos da mente humana. Talvez com o desenvolvimento dos computadores quânticos e da engenharia avançada, além da neurociência poderemos no futuro atribuir estados mentais a um software ou a um andróide, mas isso dependerá de como ambas as teorias e ciências evoluírem.

* * *

Referências

CHURCHLAND, P. S. Uma máquina poderia pensar? Trad. Nara Ebres Bachinski. São Paulo: **Cognitio-Estudos** 12, v.1, 157-169. 2015.

FERNANDES, Anita M. da R. **Inteligência artificial**: noções gerais. Florianópolis: VisualBooks, 2005.

FODOR, J. Searle sobre o que só os cérebros podem fazer. In: L. Bonjour; A. Baker (Org.) **Filosofia**: textos fundamentais comentados. São Paulo: Artmed, 2010. p. 240-242.

FRANKISH, K.; RAMSEY, W. **The Cambridge handbook of artificial intelligence**. Cambridge: Cambridge University Press, 2014.

McCULLOCH, Warren; PITTS, W. A logical calculus of ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics** 5.4(1943): 115-133.

MINSKY, M.; PAPERT, S. **An introduction to computational geometry**. Cambridge, Mass.: MIT, 1969.

PIGLIUCCI, M. **Can machines think?** Artificial intelligence & philosophy of mind.Disponível em <https://youtu.be/t7vG2WYuWks>. Acesso em 14 de out. 2015.

ROSENBLATT, F. The Perceptron: A Probabilistic Model For Information Storage And Organization In The Brain". In: **Psychological Review**. 65 (1958): 386–408.

RUMELHART, D.E; McCLELLAND, J. **Parallel distributed processing:** explorations in the microstructure of cognition. Cambridge, Mass.: MIT, 1986.

SEARLE, J. A mente do cérebro é um programa de computador? In: L. Bonjour; A. Baker (Org.) **Filosofia**: textos fundamentais comentados. São Paulo: Artmed, 2010. p. 232-239.

STARZYK, Janusz A.; PRASAD, Dilip K. A computational model of machine consciousness**. International Journal of Machine Consciousness** 3.2 (2011): 255-281.

TURING, A. Maquinário computacional e inteligência. In: L. Bonjour; A. Baker (Org.) **Filosofia**: textos fundamentais comentados. São Paulo: Artmed, 2010. p. 227-231.