XIV Semana Acadêmica do Curso de Filosofia da FAPAS
I Encontro Peter Strawson | O MÉTODO FILOSÓFICO E A METAFÍSICA
07 a 09 de outubro de 2015
ISSN 2359-6597

REDES NEURAIS ARTIFICIAIS PODEM TORNAR-SE CONSCIENTES?

Nara Ebres Bachinski*

Resumo: O funcionalismo[1] como teoria da mente teve um salto emergencial conforme o desenvolvimento da inteligência artificial (IA) se tornou eminente. A partir dessa teoria, estados mentais passam a ser caracterizados devido ao seu papel funcional em analogia a um computador. Nessa analogia os estados mentais são caracterizados em termos de *input* (entrada) e *output* (saída) de informações de um sistema ou organismo. Com o passar dos anos e o desenvolvimento da engenharia de computação e o desenvolvimento da neurofisiologia surgiram inúmeras questões acerca da IA e novas teorias da mente surgiram, tais como o conexionismo e a teoria computacional da mente. Dentre elas, a discussão da possibilidade de emergência de consciência em máquinas através das redes neurais artificiais. Essa questão vem sendo bastante debatida a partir dos anos 90.

Palavras-chave: Inteligência artificial. Consciência. Redes neurais artificiais.

Introdução

Em meados do século XX, com o desenvolvimento dos primeiros computadores, a pergunta sobre se uma máquina pode pensar feita por cientistas e ganhou novos contornos. A possibilidade de seres humanos efetivamente criarem uma máquina pensante apareceu então como algo provável para diversos pensadores. Alan Turing tratou do tema em uma série de artigos influentes. A palavra "pensamento", ele sustentou, não é adequada para uma investigação rigorosa, dada a sua ambiguidade. Usamos essa palavra para nos referirmos a atividades e fenômenos bem distintos, como, por exemplo, fazer um cálculo, pensar no que se vai fazer amanhã, lembrar-se do que se almoçou, imaginar uma situação possível, sentir-se feliz, saborear alguma comida, sentir-se satisfeito, perceber-se como indivíduo, ou seja, estados mentais de variados tipos. Não obstante, problemas semelhantes ocorrem com as palavras "consciência" e "inteligência". Em um sentido trivial de "inteligência" pode ser entendida de várias maneiras, um exemplo é a capacidade de fazer cálculos. Se entendermos inteligência como a capacidade de realizar cálculos, computadores digitais podem ser

* Graduanda Curso de Filosofia Licenciatura Plena da Universidade Federal de Santa Maria. Bolsista PIBIC/Cnpq. Email: naraebresb@gmail.com

[1] Teoria da mente segundo a qual estados mentais podem ser descritos e desempenhados por sistemas análogos, ou seja, um cérebro ou coração pode ser substituído por sistemas que desempenhem as mesmas funções.

considerados inteligentes, uma vez que eles operam cálculos e equações com muita rapidez e destreza. Quando se trata da palavra "consciência" não se tem um conceito delimitado a seu respeito e muitas vezes ela é confundida com "conhecimento". A palavra "consciência" possui inúmeras definições, a saber, a visão da psicologia, da psicanálise e da filosofia. Muitos estudiosos utilizam o conceito de "consciência" para designar estados de autoconsciência ou para se referir a qualquer estado ou forma de comportamento inteligente. Ainda não existe um consenso em relação ao conceito de consciência, esta é uma questão que ainda se encontra em aberto. No decorrer dos anos e com o avanço das tecnologias a pergunta sobre se uma máquina poderia pensar foi reformulada e passou a questão da possibilidade de máquinas tornarem-se conscientes. Essa questão permitiu o desenvolvimento de modelos de softwares baseados em redes neurais artificiais. Essas são modelos de aprendizagem baseados em redes neurais biológicas, ou seja, são sistemas interligados que trocam mensagens entre si. Assim como as redes neurais biológicas, redes artificiais também possuem "neurônios". Esses "neurônios" são baseados no modelo biológico, ou seja, possuem conexões de entrada e de saída, além dessas conexões podem ser ajustadas de acordo com a experiência, o que a torna adaptável aos inputs e também capaz de aprendizagem. No presente trabalho pretendo fazer uma recapitulação de como surgiu o problema e depois desenvolver a discussão acerca de máquinas tornarem-se conscientes por meio de redes neurais artificiais.

1 Inteligência Artificial e Redes Neurais Artificiais

Os primeiros modelos de redes neurais artificiais surgiram na década de 1940 e em 1951 Frank Rosenblatt criou a primeira rede neural artificial com o nome de *Percepton*. No Período entre 1970 até 1980 a pesquisa foi abandonada. Foi a partir da década de 1980 que avanços na capacidade de processamento dos computadores permitiram a simulação efetiva de redes neurais (RUMELHART, 1986). A partir dos anos 2000, com o desenvolvimento da neurociência, da engenharia da computação e robótica os modelos de redes neurais artificiais baseados na arquitetura do cérebro humano tiveram um maior destaque, sendo focado em modelos de redes neurais modulares. Essa arquitetura é formada por módulos distribuídos em camadas, cada um dos quais com funções específicas e interdependentes. Ou seja, nesse modelo há subdivisões cada qual com suas especificações, sendo elas subdivididas em outros módulos tais como: memória, atenção, atenção de comutação, memória episódica, etc. Essas por sua vez, são conectadas tanto hierarquicamente quanto paralelamente. Um modelo desse tipo foi esboçado por Patrícia Churchland, em *Poderia uma máquina pensar* (1990). Mais

tarde, já nos anos 2000, Janusz A. Strarzyk e Dillip K. Prasad formulam um modelo para máquinas conscientes, *A computational model of machine consciousness* (2010).

O problema da emergência de consciência em máquinas é antigo. Em *Computação e inteligência* (1950), Alan Turing levanta a questão sobre se máquinas podem pensar. A palavra "pensamento", ele sustenta, é demasiadamente vaga para permitir uma resposta clara. Mas, segundo ele, podemos tratar de questões parecidas, ou de questões que podem lançar alguma luz sobre o tema. Turing faz isso por meio de um teste, que introduz com o que chama de "jogo da imitação". Nesse jogo, há três indivíduos, um homem, uma mulher e um interrogador, cada qual em uma sala isolada, sem comunicação direta. O interrogador tenta descobrir em qual sala está o homem e em qual sala está à mulher por meio de perguntas formuladas por escrito e datilografadas e entregues ao interrogador, que as repassa aos outros dois participantes. 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 primeiro. Turing então pergunta o que aconteceria se o homem fosse substituído por um computador. Este seria capaz de enganar o interrogador com tanto sucesso quanto um homem geralmente consegue? Essas questões, Turing sugere, podem substituir a pergunta original sobre se máquinas poderiam pensar. Se for 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. Turing afirmou (1950, p. 13) que provavelmente em 50 anos, isto é, por volta do ano 2000, computadores já seriam suficientemente eficazes no jogo da imitação. Embora os computadores de hoje sejam muito mais desenvolvidos que os da época, ainda há controvérsias se passam no teste de Turing. Com certeza, temos hoje programas de computador bem mais eficazes que humanos em jogos de xadrez, por exemplo. No entanto, para o teste de Turing, não há certeza que são suficientemente eficazes na imitação do comportamento humano. Em parte isso depende de como o teste é realizado. São variáveis relevantes a duração do teste (por quanto tempo o computador tem de ser capaz de enganar seu interlocutor), que tipos de tarefas tem de ser capaz de desempenhar, qual a complexidade das respostas.

Diversos filósofos, no entanto, objetam que esse critério puramente comportamental não é suficiente para se atribuir a máquinas a capacidade de pensar. Com certeza, computadores são capazes de calcular. Entretanto, pensar exigiria algo mais. Um dos filósofos que se destacou por argumentar nesse sentido foi John Searle. Em *Mentes, cérebros e programas* (1981), John Searle apresenta o experimento mental do *quarto chinês*. Seu

objetivo é mostrar que a manipulação formal de símbolos, por si só, não produz estados mentais. Esse experimento é uma objeção ao *teste de Turing*. O *quarto chinês* consiste em um experimento mental em que um homem se encontra em um quarto onde há duas aberturas, uma de entrada e outra de saída. É oferecido ao homem textos em chinês e junto com eles um manual também em chinês, mas com as regras em inglês (língua nativa do homem). É feito ao homem questões em chinês, que devem ser respondidas também em chinês. A partir do manual e da leitura das regras em inglês o homem responde satisfatoriamente às questões, mesmo não compreendendo nada de chinês. O homem, nesse caso, está somente manipulando formalmente símbolos, mas não tem ideia do que significam. Nesse sentido, Searle argumentou, esse homem está fazendo o mesmo que um computador, uma vez que o computador possui um input determinado e fornece um output como resposta. Para Searle, a instanciação de um programa, a saber, a manipulação formal de símbolos, não é suficiente para produzir estados mentais. Assim, Searle diz que os computadores possuem apenas sintaxe e não semântica. Desse modo, sua crítica a tentativa de emergência de pensamentos a partir de modelos não orgânicos se torna clara, uma vez que ele defende a não existência de semântica em manipulações simbólicas em computadores digitais.

Outro filósofo a concordar, em partes, com Searle foi Jerry Fodor. Ele respondeu a Searle concordando que uma instanciação de um programa não é suficiente para que haja estados mentais, ou seja, a mera manipulação de símbolos não constitui um estado mental (SEARLE, 2010). Para ele também é necessário que os símbolos façam referência a algo no mundo. Contudo, 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. Segundo Fodor, se o experimento do quarto chinês fosse pensado como um software e a ele se ligasse um robô com características perceptuais, então o robô teria as relações causais do tipo adequado, pois a manipulação de símbolos estaria causalmente conectada a objetos no mundo.

É razoável supor, segundo Fodor, que o tipo certo de relação causal é a 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 ao nosso seja condição necessária para que haja esse tipo de relação. Tampouco não 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 os 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. Em outras palavras, estariam em relações causais do tipo adequado, logo poderia se atribuir estados mentais ao robô.

Patrícia e Paul Churchland, no artigo *Poderia uma máquina pensar?* (1990), objetam ao experimento mental do *quarto chinês*. Eles analisam o argumento de Searle que consiste em:

> Axioma 1: programas de computadores são formais (sintaxe).
> Axioma 2: mentes humanas têm conteúdo mental (semântica).
> Axioma 3: a sintaxe não é constitutiva nem suficiente para semântica.
> Conclusão: programas de computador não são nem constitutivos de nem suficientes para mentes (2015, p.6).

Eles sustentam que Searle comete petição de princípio, uma vez que quando analisada a premissa segundo a qual a "sintaxe por si só não é constitutiva de, nem suficiente para mentes" (CHURCHLAND, 1990, p. 162). Nesse caso, segundo eles, fica clara a petição de princípio, quando a terceira premissa é comparada a sua conclusão "programas de computador não são nem constitutivos de nem suficientes para mentes" (ibid.). Os Churchlands concordam que a instanciação de um sistema como o do quarto chinês é insuficiente para produzir fenômenos semânticos. Entretanto, disso não se segue que nunca ocorrerá de que programas de computador mais complexos não possam produzir produtos semânticos. Eles negam também que máquinas de Turing possam produzir estados mentais. Contudo, eles argumentam que computadores que imitem a arquitetura das redes neurais biológicas podem vir a se tornar conscientes. Um modelo de redes neurais consiste em uma rede que contêm várias camadas, cada uma delas com unidades conectadas em paralelo a 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 *input* (padrão de ativação), esses vetores serão convertidos em um só vetor de *output*. Esse processo é denominado por eles como "treinando a rede".

Alguns modelos mais sofisticados de redes neurais artificiais surgiram nos últimos anos. Um desses modelos será apresentado abaixo, ele é baseado em arquiteturas modulares. Essas são divididas em três módulos com funções específicas e dependentes uma da outra. Esse modelo foi apresentado por Janusz A. Strarzyk e Dillip K. Prasad em *A computational model of machine consciousness* (2010). Nesse artigo, afirmam que apesar do grande interesse de pesquisadores a respeito da consciência e sua efetuação em máquinas, a modelagem computacional de consciência ainda é muito limitada. Segundo eles, há inúmeras dificuldades,

sendo elas de ordem técnicas, filosóficas e computacionais. A principal é a dificuldade de compreensão de noções abstratas relacionadas, como pensamento, atenção, consciência, etc. O maior problema concentra-se em uma definição física de *consciência*, uma vez que não se tem uma teoria adequada que possa se tornar como base para os modelos computacionais de máquinas conscientes. Eles entendem que a inteligência e a consciência são propriedades de uma mente física e não um fenômeno metafísico.

Várias arquiteturas modulares foram propostas, essas consistem em um modelo hierárquico de processos no qual a informação flui através de módulos funcionais. Contudo, os sistemas conscientes podem não ser totalmente hierárquicos, uma vez que cada módulo funcional deve se relacionar com outros módulos em fluxos bidirecionais de informação, com um lateral correspondente, bem como o fluxo hierárquico.

O ponto de vista dos autores acerca da consciência emergente em máquinas é de que ela é vinculada à inteligência. Não obstante, a inteligência é ligada a capacidade de produzir representações sensoriais estáveis e antecipar os resultados de suas ações no meio ambiente. Para eles, um sistema que não é inteligente não pode ser consciente. Para que um sistema seja consciente é necessário que ele esteja ciente do que acontece com ele e do que acontece à sua volta. Logo, a definição de consciência utilizada pelo modelo apresentado consiste em que ela é uma função incorporada e emergente de inteligência da máquina. Não obstante, essa emergência ocorrerá e dependerá da percepção, habilidades motoras, estímulos, pensamentos e planos. Para que se tenha essa interação com o meio, segundo Strarzyk e Prasad, é necessário que haja uma inteligência corporificada, ou seja, um corpo mecânico equipado com 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 consciente de outros eventos e objetos. A máquina necessitará, para que seu aprendizado seja eficiente, de um mecanismo de *atenção* e *atenção de comutação*. "O termo *atenção* é um processo seletivo de percepção cognitiva, tomada de decisão, controle de ação ou outras experiências cognitivas"[2] (2010). E a "*atenção de comutação* corresponde a um processo dinâmico resultante da concorrência entre as representações relacionadas com motivações, inputs sensoriais e pensamentos internos, incluindo sinais espúrios. Assim, a atenção de comutação pode ser resultado de uma experiência cognitiva deliberada (e sinal assim plenamente consciente baseada no conhecimento, previsão, associação ou memória episódica) ou pode resultar de processo subconsciente (estimulada por sinais internos ou

[2] "*Attention* is a selective process of cognitive perception, decision making, action control or order cognitive experiences." (2010, p. 8).

externos). Assim, a *atenção* é uma experiência consciente, a *atenção de comutação* não tem que ser"[3] (2010). Não obstante, "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 impulsionado pela motivação da máquina e a seleção de metas, a atenção de comutação, a 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"[4] (2010). Segundo eles, é a partir do executivo central relacionando experiências cognitivas de motivações e planos que irá surgir à autoconsciência em uma máquina.

Ainda não há um consentimento entre os estudiosos de IA acerca do desenvolvimento de consciência através de redes neurais artificiais. É uma questão que continua em aberto desde 1940, com a evolução da engenharia e hardware e de software, e ainda da neurofisiologia do cérebro humano muito ainda está para ser visto. Será que realmente que em um futuro teremos máquinas inteligentes e conscientes como seres humanos? Ou há em nós algo que não possibilite esse feito? Essas questões também continuam em aberto.

Conclusão

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, visto que esse ramo da filosofia da mente, ainda está em expansão. Com o avanço da inteligência artificial, da neurociência e da engenharia algumas das questões podem vir a serem respondidas e outras ainda irão surgir ao longo das pesquisas.

[3] "*Attention switching* is a dynamic process resulting from conception between representations related to motivations, sensory inputs and internal thoughts including spurious signals. Thus attention switching may be a result of deliberate cognitive experience (and thus fully conscious signal based on knowledge, prediction, association or episodic memory) or it may result from subconscious process (stimulated by internal signals). Thus, while paying attention is a conscious experience, switching attention does not have to be." (2010, p. 8).

[4] "A machine is conscious if besides the required mechanism for perception, action, learning and associative memory, it has a central all the process (conscious or subconscious) of the machine, the central executive is driven by the machine's motivation and goal selection, attention switching, semantic and episodic memory and uses cognitive perception and cognitive understanding of motivations, thoughts, or plans to control learning, attention, motivations, and monitor actions." (2010, p. 9)

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. Poderia uma máquina pensar? Tradução de Nara Ebres Bachinski, **Cognitio Estudos**, São Paulo, v. 12, n. 1, p. 157-169, 2015.

FODOR, J. Searle sobre o que só os cérebros podem fazer. In: BONJOUR, L.; BAKER, A. (Org.) **Filosofia:** textos fundamentais comentados. 2. ed. São Paulo: Artmed Editora S.A.,2010. p. 240-242.

TURING, A. Maquinário computacional e inteligência. In: BONJOUR, L.; BAKER, A. (Org.) **Filosofia:** textos fundamentais comentados. 2. ed. São Paulo: Artmed Editora S.A., 2010., p. 227-231.

SEARLE, J. A Mente do cérebro é um programa de computador? . In: BONJOUR, L.; BAKER, A. (Org.) **Filosofia:** textos fundamentais comentados. 2. ed. São Paulo: Artmed Editora S.A., 2010. p. 232-239.

STARZYK, Janusz A.; PRASAD, Dilip K. **A computational model of machine consciousness.** Singapura: World Scientific Publishing Company, 2010.