INFORMÁTICA NA EDUCAÇÃO:
teoria & prática Porto Alegre, v.9, n.1, jan./jun. 2006. ISSN 1516-084X

Uma investigação filosófica sobre a Inteligência Artificial

Leonardo Sartori Porto

A philosophical inquiry on Artificial Intelligence

RESUMO: O objetivo deste artigo é investigar o conceito de inteligência artificial à luz da filosofia da mente contemporânea. Na primeira parte do artigo, analisaremos o argumento da Sala Chinesa de Searle e as respostas que suscitou. Na segunda parte, iremos abordar o conceito de inteligência artificial na perspectiva da relação corpo e mente.

Palavras-chave: Inteligência Artificial. Filosofia da Mente. Linguagem. Corpo e Mente.

ABSTRACT: The objective of this article is to investigate the concept of artificial intelligence to the light of the contemporary philosophy of the mind. In the first part of the article, we will analyze the argument of the Chinese Room of Searle and the answers that it excited. In the second part, we will go to approach the concept of artificial intelligence in the perspective of the relation of body and mind.

Keyword: Artificial intelligence. Philosophy of the Mind. Language. Body and Mind.

PORTO, Leonardo Sartori. Uma investigação filosófica sobre a Inteligência Artificial. *Informática na Educação: teoria & prática*, Porto Alegre, v.8, n.2, p.11-26, jan./jun. 2006.

A utilização de computadores na educação não se resume à substituição do livro impresso pelo livro digital ou ao uso de páginas de *internet* e correio eletrônico em cursos à distância: um dos principais objetivos da pesquisa em informática na educação é o de criar ambientes de ensino virtuais que possam ter um alto grau de interação com os estudantes. Para atingir este objetivo, parece ser necessário a utilização da inteligência artificial (Cumming, 1998; Wasson, 1997; Lawer, 1996), em especial no que tange à criação de tutores inteligentes (Cheung, 2003; Chou, 2003; Siemer, 1998), e à produção de *softwares* educacionais mais interativos (Brna, 1999).

O conceito de *inteligência artificial* é bastante claro e preciso nas ciências da computação; contudo, possui implicações que vão além de seu emprego técnico, pois aponta para a possibilidade de se criar pensamento nas máquinas, ou seja, a mente artificial. Aqui chegamos ao ponto focal de nosso artigo: o que é pensamento?

O filósofo John Searle, que escreveu vários textos sobre o assunto, demonstra o quanto é controverso o tema ao citar o teórico que cunhou o termo "inteligência artificial", John McCarthy: "O meu termostato tem três crenças - está demasiado quente aqui, está demasiado frio aqui e está bem aqui" (Searle, 1997a, p. 38) - se, ter crenças é pensar; então, o termostato pensa?

Searle (1990) faz uma distinção entre dois projetos distintos de inteligência artificial: o projeto "fraco", onde o computador é visto apenas como uma ferramenta que pode imitar alguns comportamentos racionais; e o "forte", cujo objetivo é produzir uma mente artificial a partir do computador. Podemos enquadrar o uso da inteligência artificial na confecção de programas de computador mais interativos e de tutores inteligentes na primeira categoria, o que diminuiria em muito as restrições que podemos ter quanto ao sentido do conceito de *inteligência artificial*, visto que este se transformaria numa metáfora. O problema é que, ao usarmos uma metáfora de maneira intuitiva e rotineira, podemos tomá-la por aquilo que ela não é: a representação fiel de algo real. Além do mais, a inteligência artificial aplicada à educação tem nos seus horizontes o objetivo de transformar o programa de computador num professor virtual[1]; raramente é dito que o "professor virtual" fará mais do que imitar, com muitas limitações, o professor humano, mas é inevitável ir mais além e perguntar: qual é a diferença entre a mente virtual e a mente real? Qual a distância que existe entre *imitar* o pensamento e pensar? Qual a diferença entre um computador que imita o enxadrista humano e o jogador de xadrez de carne e osso? Não é verdade que também o enxadrista humano aprende a jogar imitando os outros jogadores? Nós não aprendemos a falar imitando outros seres humanos? E se um computador fizesse o mesmo, porque não podemos dizer que ele está falando? E se falar com sentido implica pensar, então porque não dizer que a máquina está pensando?

Todas estas perguntas foram pensadas (e respondidas) pelo idealizador da inteligência artificial, Allan Turing, no texto "Computing Machinery and Intelligence". O seminal artigo foi publicado numa revista de filosofia (*MInd*), o que, por si só, indica que a filosofia faz parte do caminho argumentativo percorrido por essa temática.

O objetivo de nosso artigo é traçar um breve mapa do percurso filosófico da discussão sobre a inteligência artificial, tendo em vista que a filosofia pode contribuir para uma me-

lhor compreensão de tudo o que está implicado neste conceito.

Iniciaremos pelo princípio: a proposta de Turing. A seguir veremos a réplica de Searle, e a tréplica de Daniel Dennett. Desse modo é possível visualizar uma parte central do debate filosófico em torno a esse tema.

A segunda parte do artigo é dedicada a um pressuposto do projeto de inteligência artificial que raramente é levado em conta: a relação entre mente e matéria. O projeto de construção de uma mente artificial pressupõe que exista uma relação intrínseca entre a mente e a matéria, em outras palavras, o cérebro produz a mente; logo, um cérebro artificial produzirá a mente artificial. Veremos que a relação entre mente e corpo não é tão simples quanto parece, porquanto existem bons argumentos contra este pressuposto, ou seja, contra o materialismo reducionista – a teoria segundo a qual a mente pode ser inteiramente reduzida ao corpo.

O primeiro tipo de argumento dirigido contra esta doutrina vem da concepção dualista, que defende a tese da existência de um abismo intransponível entre mente e matéria. Não iremos abordar o dualismo clássico (cartesiano), onde a mente é uma substância, mas o dualismo contemporâneo, que não nega a produção da mente pelo cérebro, mas afirma que disto não se segue que os eventos mentais possam ser reduzidos a eventos físicos.

Apresentaremos, em seqüência, a réplica epistemológica do materialismo reducionista e a tréplica nomológica elaborada pelos defensores do materialismo não-reducionista.

O intuito de nosso artigo não é oferecer uma resposta definitiva ao problema, mas mostrar a sua complexidade, a fim de contribuir para a discussão teórica em torno deste assunto.

1. Prova Empírica

Se quisermos saber a resposta à pergunta “computadores podem pensar?” o melhor meio de respondê-la seria construir um computador que realizasse tal tarefa. Essa é uma típica solução empírica para resolver o problema. O idealizador dos computadores e primeiro teórico a sugerir a possibilidade de se criar uma inteligência artificial, Alan Turing (1990), propôs o seguinte teste: algumas pessoas, os juízes, farão perguntas, via terminal de computador, a uma entidade que eles não sabem se é um computador ou uma pessoa. Numa outra sala estarão pessoas conectadas a uma parte dos terminais de computador com os quais os juízes irão dialogar, a outra parte dos terminais será conectada a computadores. Caberá aos juízes determinar se estão dialogando com uma pessoa ou com uma máquina, e a tarefa dos computadores é confundir os juízes para que pensem que estão falando com uma pessoa, ao invés de uma máquina. Segundo Turing, se um computador conseguir imitar o comportamento verbal humano a ponto de confundir os juízes, então a máquina estará pensando.

A questão é saber se imitar de modo perfeito o pensamento não é pensar. Ou seja, qual a diferença entre pensar como um ser humano e, simplesmente, pensar? Turing supõe não haver nenhuma diferença e, assim, acredita que um computador que passe pelo seu teste, que seja confundido com um ser humano, realmente é uma máquina pensante.

O teste elaborado por Turing causa duas impressões imediatas: em primeiro lugar, é de grande simplicidade, mas parece ser

evidentemente implausível - de que eu não possa, baseado apenas num jogo de perguntas e respostas, distingüir as respostas dadas por uma máquina das respostas dadas por um ser humano, não se segue que a máquina pense; no mínimo, seriam necessárias algumas premissas adicionais para se chegar a esta conclusão.

Turing, é claro, antecipou as dificuldades de aceitação do seu teste, pois já existiam objeções à idéia de que as máquinas poderiam pensar, e ele reproduziu ou imaginou nove objeções, dando sua resposta a cada uma delas. Não vamos nos deter em cada uma dessas críticas e respostas, apenas apresentaremos aquelas que se referem ao teste em si.

A primeira crítica diz respeito às emoções: é apenas quando uma máquina tiver os mesmos sentimentos que os seres humanos têm ao produzirem um poema, por exemplo, que ela realmente pensará como nós. A resposta de Turing não é direta, ele primeiro nos lembra que usamos a prova oral com a finalidade de saber se um estudante realmente escreveu o trabalho que apresentou para a banca examinadora (como no caso do mestrado e do doutorado), e fazemos isso através de perguntas que explorem o sentido e o conteúdo semântico das palavras que estão envolvidas neste assunto. Turing afirma que, caso um computador se comporte do mesmo jeito, não há porque supor que não esteja se comportando do mesmo modo que os seres humanos se comportam quando realizam essa prova.

Existe outra objeção que ele nomeia como o "Argumento a partir da informalidade do comportamento". A objeção consiste basicamente na constatação de que "não é possível produzir um conjunto de regras que possa descrever o que um homem deveria fazer em cada circunstância concebível" (Turing, 1990, p. 58); e acrescenta: "disto se conclui que nós não podemos ser máquinas" (ibid.). A resposta de Turing é breve: também no caso dos computadores não é possível prever em detalhes o seu comportamento.

O artigo de Turing foi publicado em 1950, época em que não havia máquinas capazes de participar deste teste. É claro que, com o desenvolvimento dos computadores, surgiram máquinas e programas que podem competir no jogo da imitação. O filósofo Paul Churchland (1996) descreve a sua participação, como juiz, no teste de Turing, que, anualmente, é organizado pelo *Cambridge Center for Behavorial Studies* de Massachusetts - o *Prêmio Loebner*, que é conferido ao programa de computador que melhor imitar um ser humano.

Nesse concurso, computadores e pessoas precisam responder sobre apenas um tópico, por exemplo, futebol, e os interrogadores dão notas a cada candidato, sem saber se o candidato é homem ou máquina, no quesito de semelhança com o comportamento humano. Os interrogadores ficam numa sala, diante de vários terminais de computadores e cada um fará perguntas para todos os terminais, com o tempo de 15 minutos por terminal. Numa outra sala estão terminais onde há computadores e seres humanos respondendo às perguntas. Em cada sala existem juízes que irão atestar se o processo está sendo feito dentro das regras. Churchland foi um destes juízes e diz que computadores já conseguiram enganar os interrogadores, mas que, quando foi juiz, nenhum dos interrogadores foi enganado pelas máquinas; entretanto, cinco, dos oito interrogadores, tomaram um ser humano por computador!

Turing, no seu artigo, afirma ser muito fácil um ser humano imitar um computador, mas Churchland diz que este é um dos proble-

mas do teste: seres humanos não são muito hábeis, dentro das condições do teste, em distinguir outros seres humanos de máquinas.

2. Sala Chinesa

As críticas que o teste de Turing recebeu não referem, obviamente, a sua exeqüibilidade empírica – o experimento é perfeitamente realizável –, mas ao que é pressuposto pelo argumento: que basta uma máquina ter um comportamento visível idêntico ao de um ser humano quando este exerce uma atividade que requer pensamento (falar) para que se constate que a máquina pensa. Um filósofo, John Searle (1990), propôs um *experimento mental* para refutar o tipo de experimento empírico proposto por Turing[2]. O experimento ficou conhecido como o "argumento da sala chinesa" e talvez seja o experimento mental mais famoso em toda a discussão sobre a possibilidade da Inteligência Artificial.

No experimento, o próprio Searle, que desconhece o idioma chinês, está trancado numa sala onde recebe uma folha com um grupo de caracteres em chinês, depois lhe é dado uma segunda folha também com caracteres em chinês, acompanhada de outro papel com regras em inglês (língua que ele entende) para relacionar os *símbolos* da segunda folha com os da primeira. As regras escritas em inglês lhe informam que toda vez que na primeira linha da primeira folha ocorre determinado grupo de símbolos e na primeira linha da segunda folha, outro determinado grupo de símbolos, ele deverá escrever, numa terceira folha, um outro grupo de símbolos, tudo em chinês, e deverá passar esta terceira folha por uma janela para alguém do lado de fora - note que ele irá identificar os símbolos exclusivamente pelos seus desenhos, ou seja, pela sua *forma*. No experimento mental, Searle recebe as folhas em chinês por uma janela e entrega a folha que escreveu, por outra. Para quem está do lado de fora da sala, as folhas que são "inseridas" nesta contém um texto em chinês e perguntas sobre este texto, e a folha que "Searle dentro da sala" retorna ao exterior que contém as respostas às perguntas, de tal forma que quem souber chinês entenderá perfeitamente estas respostas. Para quem sabe chinês, a sala estará falando chinês, já que pode responder perguntas que lhe são feitas; porém, e Searle, sabe chinês? É claro que não, pois tudo o que está fazendo é manipular símbolos cujo significado desconhece.

Searle utiliza o experimento mental para demonstrar que, mesmo no caso do experimento empírico proposto por Turing dar o resultado esperado, não provará que os computadores estão realmente se comunicando e, portanto, que estão de fato pensando, porque tudo o que o computador fará é seguir as regras formais do seu programa, onde símbolos são substituídos por outros símbolos, mas disto não se segue que compreenda o *significado* dos símbolos. Nas palavras de Searle:

> Essencial à nossa concepção de um computador digital é que as suas operações possam ser especificadas em termos puramente formais (...) Mas esta característica dos programas, que se definem em termos puramente formais ou sintáticos, é fatal para a concepção de que os processos mentais e os processos de programa são idênticos. (...) A razão por que nenhum programa de computador pode alguma vez ser uma mente é simplesmente porque um programa de computador é apenas sintático, e as mentes são mais do que sintáticas. As mentes são semânticas, no sentido de que possuem mais do que uma estrutura formal, têm um conteúdo. (Searle, 1997, p. 38-9)

O argumento do filósofo tem como pressuposto fundamental a concepção da mente humana – e, *a fortiori*, do pensamento - como aquilo que, além de ter habilidades *sintáticas*, necessariamente possui habilidades

semânticas. Outro pressuposto é o de que as habilidades semânticas não podem ser derivadas (ou, pelo menos, *exclusivamente* derivadas) de habilidades sintáticas. Neste pressuposto está envolvido o conceito de *intencionalidade*: nossas habilidades semânticas envolvem necessariamente a intencionalidade e "(...)nenhum modelo puramente formal jamais será suficiente por si mesmo para a intencionalidade porque as propriedades formais não são por si mesmas constitutivas da intencionalidade(...)" (Searle, 1990, p. 82). O terceiro pressuposto é o de que os computadores realizam atividades que são puramente formais, ou, como ele escreve, podem ser especificadas em termos puramente formais.

O argumento da sala chinesa recebeu (e ainda recebe) inúmeras críticas, o próprio Searle (1990), imitando Turing, compilou e respondeu a estas críticas. Não vamos analisar as críticas e as réplicas de Searle, iremos apenas nos deter no que diz respeito ao segundo pressuposto, aquele que envolve o conceito de "intencionalidade". É importante ressaltar, no entanto, que, até onde sabemos, nenhuma crítica foi feita ao primeiro pressuposto, aquele que afirma que o pensamento envolve necessariamente a capacidade semântica. Muitas críticas, contudo, foram feitas ao terceiro pressuposto, visto que muitos teóricos julgam que os programas de computadores não realizam atividades puramente formais[3].

3. Intencionalidade

"Intencionalidade" é um conceito medieval que foi redescoberto pelo filósofo e psicólogo alemão do século XIX, Franz Brentano. O termo vem do verbo latino "intendo" que significa "apontar", "indicar". Brentano utilizou o conceito para definir algo que é típica e exclusivamente mental: apenas os fenômenos mentais indicam alguma coisa, representam algo, o que não ocorre com os fenômenos físicos. Imagine um pedaço de chumbo, ele não representa nada em si mesmo. Imagine, agora, que uma pessoa veja este pedaço de chumbo e perceba que há uma palavra escrita nele; a mente desta pessoa irá conectar os símbolos (as letras) que estão impressas no chumbo com aquilo que significam, isto é, com aquilo que *apontam* ou *indicam*: a palavra "árvore" aponta, por assim dizer, para o objeto a que se refere - as árvores. É importante notar que é apenas no momento em que a palavra é lida por alguém que possa compreendê-la que a intencionalidade ocorre, o pedaço de chumbo com a palavra, se não é lido por ninguém, não passa de um pedaço mudo de matéria.

Na medida em que a intencionalidade é a capacidade que a mente tem de representar objetos, nossa habilidade semântica é intencional. Mas, a intencionalidade não se refere apenas a compreender o significado de uma palavra a partir do significado de outras palavras, como ocorre no dicionário, pois implica ter crenças e desejos com relação à palavra. E ter uma crença ou ter um desejo não é algo puramente formal: ter a crença de que está chovendo, é ter um *conteúdo* mental e ter certos desejos relativos a ela (por exemplo, o desejo de pegar um guarda-chuva). Portanto, não basta dotar o computador de um dicionário para que ele possua intencionalidade, uma vez que tudo o que fará é substituir um símbolo (por exemplo, as letras da palavra "chuva") por outros símbolos (as letras da expressão: "precipitação atmosférica formada por gotas d'água") – mas não existirão crenças nem desejos relacionados a estes símbolos.

Searle (1997b) apresenta um outro argumento que poderíamos chamar de "o uso

ambíguo do conceito de intencionalidade". No início do artigo mencionei a surpresa de Searle ao constatar que, para alguns teóricos, os termostatos possuem crenças; com intuito de entender este tipo de afirmação, ele criou a distinção entre "intencionalidade original" e "intencionalidade derivada". A intencionalidade *original* pode ser ilustrada pela frase: "Estou com sede", já a intencionalidade *derivada* é exemplificada pela frase "O gramado está com sede". É claro que, no segundo caso, fala-se metaforicamente, ou como o filósofo prefere, está implícito o *como se*: faz tempo que o gramado não é regado, as gramas estão amareladas, logo podemos dizer que é *como se* o gramado tivesse o desejo por água. Apenas seres humanos e animais podem desejar e é somente com relação a eles que a intencionalidade é original.

O filósofo advoga um uso ambíguo do termo "intencionalidade" quando este se refere às máquinas: neste caso, trata-se apenas da intencionalidade derivada. O programa de computador possuiria uma intencionalidade derivada da intencionalidade original do ser humano que o produziu. Mas há uma tese implícita nesta observação: a intencionalidade derivada jamais se transformará numa intencionalidade original. É uma tese óbvia se pensarmos que a intencionalidade derivada nada mais é do que uma metáfora. Assim, para que máquinas possam ser dotadas de intencionalidade é preciso eliminar a diferença entre a intencionalidade original e a intencionalidade derivada, essa é a estratégia adotada por Daniel Dennett:

> A solução para o problema da nossa intencionalidade é direta. Nós simplesmente concordamos que os artefatos representacionais (como descrições escritas e esboços) possuem intencionalidade derivada em virtude do papel que desempenham nas atividades de seus criadores. Uma lista de compras escrita em um pedaço de papel possui apenas a intencionalidade derivada que obtém do agente que a escreveu. Bem, da mesma forma a lista de compras mantida pelo mesmo agente na memória! Ela é interna [neste último caso], não externa, mas ainda é um artefato criado pelo seu cérebro e significa o que significa em razão da posição particular na economia em funcionamento das atividades internas de seu cérebro e do papel que exerce no controle das atividades complexas do seu corpo no mundo real que nos cerca. (Dennett, 1997, p. 53)

A tese de Dennett soa bastante contra-intuitiva: ele nega a diferença entre *ter um conteúdo na mente* e *ter o mesmo conteúdo fora da mente*. Para o filósofo, a intencionalidade da mente humana é derivada da evolução que criou nossos corpos, sendo o resultado do desenvolvimento de intencionalidades primitivas. Portanto, a nossa intencionalidade é uma intencionalidade derivada, e não, original, como pensa Searle. Ora, se a nossa intencionalidade é derivada e, ainda assim, nós pensamos, então por que um computador, cuja intencionalidade é obviamente derivada da nossa, não pode pensar?

É claro que um computador precisa fazer mais do que substituir símbolos para demonstrar intencionalidade, é necessário que demonstre capacidade de interagir com o mundo físico; por esta razão, Dennett propõe que o computador seja colocado num robô que possa se relacionar com o mundo, tanto no sentido de perceber o que o cerca (adaptando câmeras de TV e outros sensores no robô), quanto no de modificar o seu ambiente (construindo braços mecânicos para o robô e um meio de locomoção). Além do aspecto mecânico, é fundamental que o computador-robô não seja inteiramente programado no que diz respeito às suas ações: o robô deverá realizar certas tarefas especificadas no programa,

também terá regras que o orientarão na realização de suas tarefas, mas deverá *aprender* com o ambiente, através da tentativa e do erro, a melhor forma de realizar sua tarefa.

O nosso robô terá duas características que aproximam a sua intencionalidade derivada daquela dos seres humanos: 1) deverá interagir fisicamente com o ambiente, 2) e deverá ser capaz de aprender a partir desta interação. É interessante notar que o artigo de Turing já falava de máquinas que aprendem, e ele indica que talvez este seja o melhor caminho para produzir a inteligência artificial. A tecnologia atual já está bastante desenvolvida com relação à construção de robôs, e existe o objetivo de criar robôs que não apenas interajam com o ambiente como também aprendam através desta interação[4]. Nesse sentido, parece apenas uma questão de tempo o desenvolvimento de um robô que possa demonstrar comportamento intencional e, por conseguinte, pensamento.

Já é o momento de avaliar o esclarecimento que as reflexões filosóficas proporcionam com relação ao uso da inteligência artificial na educação. Antes de qualquer coisa, é preciso lembrar o que Wittgenstein escreveu sobre a filosofia: "A filosofia deixa tudo como está" (Wittgenstein, 1987, p. 262). A reflexão filosófica sobre a inteligência artificial não visa a estabelecer a maneira como esta pode ser utilizada na educação, tampouco é seu objetivo determinar de que modo a inteligência artificial pode ser obtida. A filosofia pode apenas iluminar as várias facetas do conceito "inteligência artificial" a fim de elucidar aquilo que está envolvido neste conceito. Ele não é uma simples metáfora, visto que um computador não está tão distante do cérebro humano quanto o sol está distante do sorriso ou arcoíris da ponte. Esta é a conclusão que podemos depreender do pensamento de Turing, Searle e Dennet: embora tenham concepções diferentes sobre a inteligência artificial, todos concordam que o cérebro é uma máquina. Sendo uma máquina, o que ele produz, a princípio, pode ser reproduzido em outra máquina – é o que ocorre quando leio o mesmo texto que outra pessoa está lendo, pois as "máquinas" que temos em nossas cabeças estão produzindo pensamentos similares.

Searle não acredita que possamos recriar a mente humana num computador, mas esta impossibilidade não parece compatível com o materialismo que ele professa. Turing, é claro, sustenta a concepção oposta, que é bastante contra-intuitiva: como imaginar que um computador possa ser um professor em sentido estrito? Este é o desafio que os defensores da inteligência artificial no sentido forte fazem ao senso comum. Não precisamos da filosofia para saber que esta possibilidade existe, mas a discussão filosófica que acabamos de acompanhar indica que o candidato mais capaz para ocupar a vaga de professor artificial não é o computador tradicional, mas uma máquina que possa interagir com o mundo e que tenha capacidade de aprender – nunca estaremos livres da tarefa de formar professores, mesmo os artificiais.

3. Mente e matéria

Começamos o nosso artigo apresentando a primeira tentativa de prova empírica de que uma máquina pode pensar. Nesta prova está envolvido um programa de computador que pode imitar a capacidade lingüística de um ser humano. Já existem programas que podem realizar esta tarefa, mas vimos, através do argumento da sala chinesa, que a prova empírica não é suficiente. Surge, então, um novo tipo de prova empírica: um computador

com corpo, um robô, que pode aprender através da interação com o meio, com os seres humanos e, quem sabe, possa se criar uma sociedade de robôs - o que replicaria todo o ambiente no qual as mentes humanas existem.

É claro que muitas críticas foram feitas a esta nova tentativa empírica de comprovar a Inteligência Artificial. O próprio Searle e mesmo Dennett teceram argumentos contra esta possibilidade[5]. Na parte restante do artigo, vamos nos dedicar a um pressuposto que engloba tanto a possibilidade do programa de computador, do robô e, mesmo, do conexionismo (a tentativa de se criar redes neurais artificiais): a tese de que a mente é criada pela matéria.

A criação de uma *mente artificial* implica a criação de um artefato que produza esta mente - seja um computador digital, um robô ou um sistema conexionista -; por trás desta idéia está o pressuposto de que a mente é o resultado de um processo físico que ocorre em nosso cérebro (ou em todo o nosso corpo) e que, como qualquer processo físico, pode ser reproduzido em um sistema físico que seja organizado de tal modo a obter este resultado. Se a mente pode ser produzida por um sistema constituído por neuroproteínas, porque não poderia ser produzida por um sistema feito de metal e silício[6]?

Em nossa atual concepção científica do mundo esta pergunta soa como puramente retórica, pois é evidente que, se a matéria, numa determinada organização, cria a mente, então basta reproduzir esta organização para se criar uma mente artificial. O problema é que desde o início da história da filosofia existem teorias que negam a identidade entre mente e matéria. Entre os primeiros filósofos encontra-se Anaxágoras cuja cosmologia afirmava que "todas as coisas estavam unidas e imóveis por um período infinito de tempo, e que a Mente introduziu movimento e as separou"[7]. Para Anaxágoras, portanto, é a mente quem ordena a matéria[8]. Já outro filósofo pré-socrático, Demócrito, identificava mente com corpo, sendo um dos fundadores do materialismo. Atomista, ele acreditava que a mente era composta por átomos, na verdade, átomos esféricos, que também constituíam o fogo e eram mais propícios para causar o movimento.

A filosofia moderna começa com a disputa entre um filósofo materialista, Thomas Hobbes, e o pai do dualismo moderno, René Descartes. Não apenas o materialismo chegou até nossos dias, existem versões contemporâneas do dualismo e concepções materialistas que negam a possibilidade de se reduzir a mente à matéria. Iremos analisar brevemente estas concepções para saber em que sentido elas estão relacionadas com o projeto da Inteligência Artificial.

4. Dualismo contemporâneo

Certamente o dualismo é uma posição minoritária na discussão filosófica contemporânea sobre a relação entre a mente e o corpo. Ainda assim, o número de teses a argumentos dualistas é muito grande para serem analisados neste artigo; em virtude disto, vamos analisar apenas um artigo que ficou bastante famoso na filosofia da mente contemporânea: "What is it like to be a bat", do filósofo norte-americano Thomas Nagel (1974).

O artigo principia questionando a tese materialista de que todo evento ou fenômeno mental pode ser *reduzido* a um evento ou fenômeno físico. Quando a ciência explica um fenômeno, ela o transforma num fenômeno científico, por exemplo: quando a ciência explica que a água é composta de dois átomos de hidrogênio e um átomo de oxigênio, ela está

"reduzindo" a substância "água" para a substância "H2O". Toda a redução que a ciência opera não pode "deixar resto", ou seja, quando reduzo "água" a "H2O" não sobra nada - a água é inteiramente H2O. O mesmo ocorre ou pode ocorrer com os fenômenos mentais? Podemos reduzi-los a fenômenos físicos sem "deixar resto"?

Nagel também recorre a um experimento mental. Ele pede que o leitor se coloque no lugar de um morcego. Como é perceber o mundo através do sonar natural dos morcegos? É óbvio que a experiência sensorial que os morcegos possuem não pode ser imaginada por nós. Agora, vamos mudar o experimento. Imaginemos extraterrestres que também percebem a luz visível, mas são incapazes de perceber as cores, como eles poderão imaginar a nossa sensação de cor? Note que a luz é um fenômeno físico, um fenômeno que pode ser percebido de um ponto de vista *externo*: tanto nós humanos, quanto animais que possuem visão e mesmo possíveis extraterrestres têm acesso a este fenômeno, em outras palavras, ele é *objetivo*. Já a percepção que temos das cores, ou a percepção auditiva do morcego, pertence a um ponto de vista *interno* e, portanto *único*: apenas os morcegos percebem as coisas do modo como os morcegos as percebem, apenas nós percebemos as cores da maneira como as percebemos.

O que o experimento visa a estabelecer é uma irredutibilidade do subjetivo ao objetivo. Nagel não nega que a experiência subjetiva, enfim, nossa consciência, não se relacione com o mundo, o que nega é a possibilidade de explicar a nossa consciência a partir de fenômenos físicos, porque esta explicação deixaria algo de fora - a redução não seria completa. Ele não está defendendo a tese cartesiana de que a mente é uma substância distinta do corpo, apenas sustenta que não podemos explicar a mente a partir do corpo.

Que lições podemos tirar do dualismo contemporâneo para o projeto da inteligência artificial? Como vimos mais acima, morcegos percebem o mundo, mas nós somos incapazes de imaginar como é perceber o mundo no modo dos morcegos. Podemos construir um robô que seja capaz de perceber o mundo, mas isto jamais nos ajudará a saber como *nós* percebemos o mundo, pois o robô irá perceber o mundo à maneira dele. Ora, como um dos objetivos da inteligência artificial é explicar, através da simulação de nosso pensamento em máquinas, a cognição humana, este objetivo parece não ser realizável, se Nagel está certo. Por outro lado, quando o dualismo estabelece uma distância intransponível entre a mente e a matéria (o que não é exatamente o caso do dualismo de Nagel, mas certamente é o caso de outros defensores contemporâneos do dualismo), então, apesar de existir uma ligação entre mente e matéria, jamais seremos capazes de perceber qual é esta ligação, o que implica a incapacidade de saber que tipo de configuração da matéria causa a mente - uma pedra de cal no projeto da Inteligência Artificial.

5. A resposta materialista

É claro que os defensores do materialismo elaboraram inúmeras respostas ao desafio dualista. Vamos apresentar a resposta oferecida por David Lewis (2000), porque apresenta um argumento muito claro e bem articulado, além de ter influenciado profundamente a discussão sobre este tema.

Lewis, em primeiro lugar, define a teoria da identidade psicofísica: "A teoria da identidade [psicofísica] é a hipótese de que – não por necessidade, mas por uma questão de fato – toda a experiência [eventos mentais] é idênti-

co com algum tipo de estado físico"(Lewis, 2000, p. 162). Em outras palavras, a teoria da identidade psicofísica afirma que a cada evento mental corresponde um evento físico. Para provar tal princípio é necessário partir de premissas que não o pressuponham, e Lewis aponta duas premissas que julga auto-evidentes. A força da primeira reside em ser uma mera análise do conceito de experiência: "A primeira das minhas duas premissas para estabelecer a teoria da identidade é o princípio de que a característica definitiva de qualquer experiência como tal é a causalidade [causal role]" (ibid., p.165). Podemos chamar esta premissa de princípio empirista: sempre que há experiência há relações causais.

A outra premissa, que julga ser uma hipótese plausível, é a de "(...)que existe um corpo unificado de teorias científicas que nós aceitamos e que fornece uma explicação completa e verdadeira de todo fenômeno físico (i.e. todo o fenômeno que pode ser descrito em termos físicos)" (Ibid., 169). A segunda premissa é a idéia de que a ciência consegue dar uma explicação satisfatória da realidade (aqui entendida como o conjunto de fenômenos físicos).

Existe, é claro, uma observação adicional: todas as nossas "experiências mentais" estão associadas a experiências físicas, ou seja, não existe mente sem corpo. Ora, pela primeira premissa temos que toda experiência deve implicar uma relação causal, assim, a relação entre a experiência mental e a experiência física deve ser uma relação causal. Como as ciências têm a capacidade de explicar as relações causais entre fenômenos físicos, nada impede que elas possam explicar as relações causais entre fenômenos físicos e mentais.

O filósofo admite que possam existir fenômenos que não sejam físicos, mas como não têm poder causal - pois, caso tivessem esse poder, seriam explicados pela ciência -, então eles não constituem experiência, uma vez que a primeira premissa afirma que toda experiência deve ter poder causal.

O materialismo propõe um dilema ao dualismo. A tese dualista é a de que os fenômenos mentais não podem ser explicados (reduzidos) pelos fenômenos físicos. Os dualistas precisam admitir, entretanto, que os fenômenos mentais possuem poderes causais, caso contrário, a mente seria inerte. O problema é que as relações causais são definidas pela maneira como ocorrem no mundo físico e cabe à ciência investigá-las. É claro que esta, em seu estágio atual, não consegue explicar completamente a relação entre os fenômenos mentais e os fenômenos físicos, mas o dualista vai além: afirma que ela nunca será capaz de explicar esta relação. Adicionando a premissa de que tudo o que pode ser experienciado é algo que admite relações causais, então sobram duas alternativas para o dualista: ou concede que os eventos mentais não sejam passíveis de relações causais, tornando-os não experienciáveis, ou admite que sejam passíveis de relações causais, mas, nesse caso, nada impede que possam ser causalmente relacionados a fenômenos físicos, a não ser que existam num outro tipo de realidade, uma espécie de realidade não-física. Postular uma realidade deste tipo acarreta o ônus de provar a sua existência.

6. Monismo anômalo

A teoria que vamos examinar não tenta dissolver o dilema apresentado mais acima, por que não defende o dualismo. Porém, apesar de não ser dualista, tampouco é materialista no sentido estrito, visto que nega a tese de que todo evento (fenômeno) mental pode ser

reduzido a um evento (fenômeno) físico. O *monismo anômalo* defende uma concepção que à primeira vista parece contraditória: 1) eventos mentais interagem casualmente com eventos físicos; 2) toda interação causal pode ser expressa por uma lei; mas, 3) não existe leis que conectem os eventos mentais aos eventos físicos e vice-versa. Segundo a teoria, "eventos mentais são idênticos a eventos físicos"(Davidson, 1989, p. 209) – por isto ela é *monista* (o oposto do dualismo) –, mas disto não se segue que eles possam ser reduzidos a eventos físicos, porque para que isto ocorra é preciso que os eventos mentais sejam subsumidos às leis que se aplicam aos eventos físicos, o que a teoria nega ser possível – daí o adjetivo "anômalo", ou seja, aquilo que não segue leis.

O monismo anômalo foi proposto pelo filósofo Donald Davidson. Não há espaço para expor em detalhe seus argumentos, assim, tudo o que podemos fazer é um esboço da teoria.

Uma lei da natureza ou lei científica que explique os eventos mentais a partir de eventos físicos é uma lei psicofísica. Davidson precisa provar que não existem leis psicofísicas. Leis se aplicam a objetos que possuem afinidade categorial: eu não posso aplicar leis civis a objetos inanimados, do mesmo modo, não posso aplicar leis físicas aos números (não faz sentido eu dizer que o número 3 é mais leve que o 4). É evidente que podemos aplicar leis físicas ao corpo humano, afinal, ele é um objeto físico, mas podemos aplicar leis físicas à mente humana?

Davidson argumenta que:

> Do mesmo modo que a satisfação das condições de medição da massa ou da extensão podem ser vistas como constitutivas do campo de aplicação das ciências que aplicam estas medições, a satisfação das condições de consistência e coerência racional podem ser vistas como constitutivas do campo de aplicação de conceitos como crença, desejo, intenção e ação. (Ibid., p. 236-7)

Em outras palavras, as categorias que a física utiliza incluem conceitos como massa, extensão e velocidade, mas não incluem conceitos referentes à racionalidade: quando explico porque fulano decidiu sair levando um guarda-chuva, apesar de não estar chovendo, uso categorias como a crença que ele tem na previsão do tempo apresentada na televisão e o desejo de não se molhar – são categorias que não tem correspondência no mundo físico. É constitutivo da explicação de nossas ações que eu possa me referir ao desejo que originou a ação e a crença que guiou este desejo (no caso do exemplo, a *ação* é pegar o guarda-chuva, o *desejo* é o de não se molhar e a *crença* é a de que hoje choverá).

Visto que as categorias que usamos para explicar os eventos mentais não são as mesmas que a física utiliza para explicar os eventos físicos, então a resposta à pergunta necessariamente é negativa: o campo categorial dos eventos físicos não é o mesmo que o dos eventos mentais; logo, não se podem aplicar leis físicas aos eventos mentais. É importante ressaltar, contudo, que isso não impede que eventos mentais sejam idênticos a eventos físicos, apenas não se pode estabelecer leis causais entre eles - assim, existe um campo do mental que é irredutível ao que é físico.

Não é tão fácil saber o que o monismo anômalo acarreta para o projeto da Inteligência Artificial. Em primeiro lugar, é uma teoria materialista, portanto defende a tese de que o mental é idêntico ao físico, uma tese que implica a possibilidade de se criar um mecanismo que possa produzir a mente. Por outro lado, ao

afirmar que as leis físicas não se aplicam ao fenômeno mental, ele impede o conhecimento de quais leis físicas causam o surgimento da mente. Assim, mesmo que possamos construir uma máquina que pense, analisar o seu funcionamento em nada nos ajudará a compreender como a mente funciona.

Conclusão

O projeto de criar uma inteligência artificial é um empreendimento empírico – afinal é o projeto de construir uma máquina que pense –, mas tem pressupostos que transcendem a investigação empírica. O primeiro pressuposto é a própria definição do que seja *pensamento*. O que está em jogo aqui é saber se uma definição meramente behaviorista esgota o sentido deste conceito. É evidente que o pensamento está ligado a ação, mas daí se segue que toda ação esteja ligada a um pensamento? No caso do teste de Turing, é verdadeiro que toda a vez que um ser humano fala ele está pensando, mas disto pode-se inferir que tudo o que fala pensa? Esta objeção também se estende ao robô, que pode demonstrar um comportamento similar ao humano e, ainda assim, não pensar.

A objeção, contudo, não tem a força lógica para eliminar a possibilidade da inteligência artificial. Se é verdade que uma máquina ter comportamento igual ao de um ser humano quando este age não implica que a máquina pense, tampouco a possibilidade de que ela pense está excluída. O que Searle e Nagel parecem estar afirmando é que esta máquina, para pensar, necessariamente deverá ter "vida interior", ou seja, possuir subjetividade. De qualquer modo, precisaremos ir além do behaviorismo para determinar se uma máquina é capaz de pensar.

Existe, contudo, um pressuposto mais primordial: a relação entre corpo e mente é intrínseca ou extrínseca? A tese dualista afirma que é extrínseca, ou seja, os predicados mentais são independentes dos predicados físicos. Disto se segue que não podemos saber se uma determinada característica (predicado) de um sistema físico corresponde a uma determinada característica de um sistema mental. O dualismo não exclui a possibilidade de um sistema físico qualquer (seja o cérebro, seja um computador ou um robô) produzir a mente, mas ele exclui a possibilidade de entendermos como o sistema físico está relacionado com a mente. Isto implica que podemos replicar o cérebro humano num outro meio físico - metais e silício - sem com isto replicarmos a mente humana.

A concepção materialista é a que mais favorece a possibilidade da criação de uma mente artificial, pois reduz o corpo à mente e, assim, se conseguirmos replicar com sucesso o cérebro humano em outro meio físico não há nada que impeça o surgimento de uma mente a partir deste meio físico.

O materialismo não-reducionista – o monismo anômalo – põe um freio no otimismo triunfante do materialismo tradicional. Segundo nossa interpretação, o monismo anômalo representa um obstáculo ao projeto de inteligência artificial similar ao do dualismo, pois cria uma barreira – neste caso, epistemológica – entre a mente e a matéria: as leis que guiam o comportamento da mente não são as mesmas que guiam o comportamento da matéria. Ora, quando construímos um "cérebro artificial" estamos seguindo as leis físicas; caso estas sejam diferentes das leis mentais, então não poderemos saber se do funcionamento físico do cérebro resultará uma mente artificial.

De modo algum, as reflexões filosóficas nos levam a um veredicto definitivo sobre o projeto de se construir uma inteligência artificial. O que elas nos oferecem é uma visão clara dos pressupostos não empíricos deste projeto, ou seja, os fundamentos, por assim dizer, metafísicos da inteligência artificial. Estes aspectos nos alertam para não confundir o adjetivo com o substantivo: quando falamos de um "programa inteligente" estamos usando "inteligência" apenas enquanto um adjetivo, cuja função semântica é distinguir o grau de complexidade de um programa, mas ainda não chegamos à inteligência propriamente dita. É claro que, como vimos no caso de, John McCarthy, esta confusão eventualmente ocorre, pois ele toma o "estado" de um termostato por uma "crença", quando os filósofos nos indicam que ter uma crença é mais do que uma mera mudança de estado. Assim, ter um comportamento inteligente é mais do que apenas realizar determinadas tarefas.

Com relação à educação, é importante estar atento a esta distinção para não confundirmos um programa bastante interativo com um professor artificial, ou um programa que imita a atividade docente, por um programa que *é* um docente. Por outro lado, tampouco podemos negar a possibilidade de que um dia tenhamos computadores ou robôs que possam substituir os professores, restando aos humanos programar estes artefatos.

Referências

ARISTÓTELES. *The Complete Works of Aristotle (editado por Jonathan Barnes).* Princeton; Princeton University Press, 1995.

BODEN, Margaret (ed.). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press, 1990.

BRNA, Paul. "Artificial intelligence in educational software: has its time come?" *British Journal of Educational Technology*, Vol 30, Nº 1, p.79-81, 1999.

CHEUNG, B. et alii. "SmartTutor: an intelligent tutoring system in web-based adult education". *Journal of Systems and Software*, Vol. 68, Nº 1, p. 11-25, 2003.

CHOU, Chih-Yueh et alii. "Redefining the learning companion: the past, present, and future of educational agents". *Computers & Education*, Vol. 40, Nº 3, p. 255-269, 2003.

CHURCHLAND, Paul. *The engine of reason, the seat of soul*. Massachusetts: MIT Press, 1996.

COLE, David. 'Thought and Thought Experiments', *Philosophical Studies,* Vol. 45, Nº: 3, p. 431-44, 1984.

CUMMING, G. "Artificial intelligence in education: an exploration". *Journal of Computer Assisted Learning*, Nº 14, p. 251-259, 1998.

DAVIDSON, Donald. *Essays on actions and events*. Oxford: Oxford University Press, 1989.

DENNETT, Daniel. The Intentional Stance. Cambridge: MIT Press, 1987.

DENNETT, Daniel. Tipos de Mente. Rio de Janeiro: Rocco, 1997.

LAWER, Robert. "Thinkable Models". *Journal of Mathematical Behavior*, Nº 15, p. 241-259, 1996.

LEWIS, David. "An argument for the identity theory, with addenda". In: ROSENTHAL, David (ed.). *Materialism and the mind-body problem*. Indianapolis: Hackett, 2000. P: 162-171.

NAGEL, Thomas. "What is it like to be a bat?" *The Philosophical Review,* Vol. LXXXIII, Nº 4, p. 435-50, 1974.

ROSENTHAL, David (ed.). *Materialsm and the mind-body problem*. Indianapolis: Hackett, 2000.

SEARLE, John. "Minds, brains, and programs". In: BODEN, Margaret (ed.). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press, 1990. P. 67-88.

SEARLE, John. *Mente, cérebro e ciência*. Lisboa: Edições 70, 1997a.

SEARLE, John. *A redescoberta da mente*. São Paulo: Martins Fontes, 1997b.

SIEMER, Julika; ANGELIDES, Marlos. "A comprehensive method for the evaluation of complete inteligent tutoring systems". *Decision Support Systems*, Nº 22, p. 85-102, 1974.

TURING, Alan. "Computing machinery and intelligence". In: BODEN, Margaret (ed.). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press, 1990. P. 40-66.

WASSON, Barbara. "Advanced educational technologies: the learning environment", *Computers in Human Behavior*, Vol. 13, Nº 4, p. 571-594, 1997.

WITTGENSTEIN, Ludwig. *Investigações Filosóficas*. Lisboa: Fundação Calouste Gulbenkian, 1987.

Notas

[1] No artigo "Redefining the learning companion: the past, present, and future of educational agents" de Chou et alii, os pesquisadores nos lembram que a aprendizagem privada - i. e. com um tutor humano - é quatro vezes mais eficiente que a aprendizagem na sala de aula tradicional. Neste sentido, um tutor eletrônico que consiga imitar um tutor humano pode ser até mais eficiente (pelos menos em alguns aspectos) que as aulas tradicionais.

[2] Na verdade, o experimento mental foi feito para refutar a proposta do cientista Roger Schank e seus colegas de criarem um programa de computador que simule a capacidade humana de ler e entender textos.

[3] Entre os vários artigos que expressam este ponto de vista, temos: o artigo de Margaret Boden, "Escaping from the Chinese Room" (in Boden, 1990), de Cole, 'Thought and Thought Experiments', Philosophical Studies, 45:431-44 , e o artigo 'Fast Thinking', que está no livro The Intentional Stance, de Daniel Dennett (Cambridge: MIT Press, 1987).

[4] Pesquisadores do laboratório Lira da Universidade de Genova criaram um robô que imita o comportamento de um bebê quando este percebe o mundo. O BabyBot foi programado com uma lista mínima de instruções a fim de que, a partir da interação com o meio exterior possa aprender a perceber os objetos.

[5] Searle critica a hipótese do robô no já mencionado artigo "Mind, Brain ans Programs" (Searle, 1990), e Dennett o faz no artigo "Cognitive Wheels: The Frame Problem of AI" (In Boden, 1990).

[6] Adaptado de um argumento de Margaret Boden na introdução do livro Philosophy of Artificial Intelligence (Boden, 1990, p. 6).

[7] Citado por Aristóteles, Física, 250b25.

[8] Na verdade, aquele que é considerado o fundador da filosofia ocidental, Tales de Mileto, acreditava que o universo estava cheio de espíritos e que mesmo a matéria que julgamos inanimada possuía mente - era desta forma que ele explicava o movimento dos magnetos.