MENTES, CÉREBROS E PROGRAMAS♠

John R. Searle

(University of California, Berkeley)

Tradução: Cléa Regina de Oliveira Ribeiro♣

Resumo: Este artigo pode ser visto como uma tentativa de explorar as consequências de duas proposições. (1) Intencionalidade em seres humanos (e em animais) é um produto de características causais do cérebro. Suponho que este seja um fato empírico sobre as reais relações entre os processos mentais e os cérebros. Afirmando simplesmente que certos processos cerebrais são suficientes para a intencionalidade. (2) Instanciar um programa de computador nunca é por si só uma condição suficiente para a intencionalidade. O principal argumento deste artigo é direcionado a estabelecer essa afirmação. A forma do argumento é para mostrar como um agente humano poderia instanciar o programa e ainda não ter a intencionalidade relevante. Essas duas proposições têm as seguintes consequências: (3) A explicação de como o cérebro produz intencionalidade não pode ser de que ele o faça instanciando um programa de computador. Esta é uma consequência lógica estrita de 1 e 2. (4) Qualquer mecanismo capaz de produzir intencionalidade deve ter poderes causais iguais aos do cérebro. Esta é para ser uma consequência trivial de 1. (5) Qualquer tentativa de literalmente criar intencionalidade artificialmente (IA forte) não poderia ter sucesso apenas projetando programas, mas teria que duplicar os poderes causais do cérebro humano. Isto decorre de 2 e 4. "Uma máquina poderia pensar?". No argumento apresentado aqui, *apenas* uma máquina poderia pensar, e apenas tipos muito especiais de máquinas, a saber, cérebros e máquinas com poderes causais internos equivalentes aos dos cérebros. E é por isso que a IA forte tem pouco a nos dizer sobre o pensamento, já que não se trata de máquinas, mas de programas, e nenhum programa por si só é suficiente para pensar.

Palavras-chave: Inteligência artificial; Cérebro; Intencionalidade; Mente.

Abstract: This article can be viewed as an attempt to explore the consequences of two propositions. (1) Intentionality in human beings (and animals) is a product of causal features of the brain I assume this is an empirical fact about the actual causal relations between mental processes and brains It says simply that certain brain processes are sufficient for intentionality. (2) Instantiating a computer program is never by itself a sufficient condition of intentionality The main argument of this paper is directed at establishing this claim The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. These two propositions have the following consequences (3) The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program. This is a strict logical consequence of 1 and 2. (4) Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1. (5) Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. This follows from 2 and 4. "Could a machine think?". On the argument advanced here only a machine

♠ Nota dos Organizadores do Dossiê: o filósofo estadunidense John Rogers Searle adquiriu notoriedade no campo da Filosofia da Mente em 1980, ano em que publica o influente artigo *Minds, Brains, and Programs* no periódico *The Behavioral and Brain Sciences* (v. 3, n. 3, p. 417-24). A tradução desse texto, realizada pela Professora Cléa Regina de Oliveira Ribeiro, foi publicada originalmente como capítulo de livro em *Cérebros, máquinas e consciência:* uma introdução à filosofia da mente (São Carlos: Editora da UFSCar, 1996, p. 61-94), organizado por João de Fernandes Teixeira. Expressamos, aqui, o nosso agradecimento ao Professor João Teixeira pela autorização da republicação dessa versão em português em nosso número especial. No intuito de adaptá-la às normas da *Revista Reflexões*, realizamos alguns ajustes no texto, a saber: estruturais (inclusão do Resumo/Palavras-chave, do Abstract/Keywords e de referências), ortográficos (substituição de "robot" por "robô" e de "hamburguers" por "hambúrgueres", exclusão dos tremas etc.) e tipográficos (acrescentamos alguns itálicos e aspas com base no texto original).

♣ Graduada em Artes Plásticas pela Fundação Armando Álvares Penteado (1976), graduada em Filosofia pela Universidade Estadual Paulista Júlio de Mesquita Filho (1984), Mestre em Filosofia pela Pontifícia Universidade Católica de São Paulo (1997) e Doutora em Saúde Pública pela Universidade de São Paulo (2002). Lattes: http://lattes.cnpq.br/2311892149976882.

could think, and only very special kinds of machines, namely brains and machines with internal causal powers equivalent to those of brains And that is why strong AI has little to tell us about thinking, since it is not about machines but about programs, and no program by itself is sufficient for thinking.

Keywords: Artificial intelligence; Brain; Intentionality; Mind.

Distingo entre Inteligência Artificial no sentido forte e no sentido fraco♣. De acordo com a IA no sentido forte, computadores adequadamente programados literalmente têm estados cognitivos, e, assim sendo, programas são teorias psicológicas. Argumento que a IA no sentido forte deve ser falsa, uma vez que um agente humano poderia instanciar um programa e, mesmo assim, não ter estados mentais. Examinarei alguns argumentos contra esta afirmação e explorarei algumas consequências do fato de que o cérebro de seres humanos e de animais são a base causal da existência de fenômenos mentais.

Que significado psicológico e filosófico devemos atribuir aos esforços feitos recentemente para simular capacidades cognitivas humanas através do computador? Para responder esta questão, considero útil distinguir entre o que denomino de IA no sentido "forte" e IA no sentido "fraco" ou "cautelosa". De acordo com a IA no sentido fraco, o principal valor do computador para o estudo da mente reside no fato de que este nos fornece uma ferramenta extremamente poderosa. Por exemplo, ele nos permite formular e testar hipóteses de maneira mais rigorosa e precisa do que antes. Mas de acordo com a IA no sentido forte, o computador não é meramente um instrumento para o estudo da mente. Muito mais do que isso, o computador adequadamente programado *é* uma mente, no sentido de que, se lhe são dados os programas corretos, pode-se dizer que eles *entendem* e que eles têm outros estados cognitivos. Conforme a IA no sentido forte, uma vez que o computador programado tem estados cognitivos, os programas não são meros instrumentos que nos capacitam a testar explicações psicológicas: os programas constituem as próprias explicações. Não tenho objeções à IA no sentido fraco, pelo menos no que diz respeito ao escopo deste artigo. Minha discussão será dirigida às afirmações que defini como caracterizando a IA no sentido forte, especificamente a ideia de que computadores adequadamente programados têm estados cognitivos e que os programas, a partir disso, explicam a capacidade cognitiva humana. Quando eu me referir à IA, estarei considerando a IA no sentido forte, definida através das duas afirmações acima.

Analisarei o trabalho de Roger Schank e seus colegas em Yale (cf. Schank & Abelson, 1977), porque estou mais familiarizado com ele do que com outros trabalhos semelhantes. Além

♣ Nota da Tradutora (N. T.): o termo "Inteligência Artificial" será abreviado, daqui em diante, por IA.

do mais, ele fornece um exemplo claro do tipo de trabalho que desejo examinar. Mas nada do que apresento a seguir depende de detalhes do programa de Schank. Os mesmos argumentos se aplicariam ao programa de Winograd (1972), SHRDLU, ao programa ELIZA de Weizenbaum (Weizenbaum, 1965), e a qualquer simulação de fenômenos mentais humanos baseada na máquina de Turing.

Deixando de lado vários detalhes, pode-se descrever o programa de Schank da seguinte maneira: seu objetivo é simular a habilidade humana de compreensão de histórias. É característico na habilidade dos seres humanos para compreender histórias que estes possam responder questões sobre elas, mesmo se a informação não estiver explicitamente dada no texto. Neste caso, por exemplo, suponha que seja fornecida a seguinte história: "Um homem foi a um restaurante e pediu um hambúrguer. Quando o hambúrguer chegou, estava torrado, e o homem furioso saiu esbravejando do restaurante sem pagar e nem deixar gorjeta". Ora, se a seguinte questão for formulada: "O homem comeu o hambúrguer?", você presumivelmente responderá: "Não, ele não comeu". Da mesma maneira, se for dada a seguinte história: "Um homem foi a um restaurante e pediu um hambúrguer; ao chegar o pedido, ficou bastante satisfeito e na hora de ir embora deu uma boa gorjeta à garçonete antes de pagar sua conta". Se a seguinte questão for formulada: "O homem comeu o hambúrguer?". Você certamente responderá: "Sim, ele comeu o hambúrguer". Ora, a máquina de Schank pode responder a questões deste tipo sobre restaurantes. Para poder fazer isto, ela tem a "representação" do tipo de informação que os seres humanos têm sobre restaurantes, o que a torna capaz de responder tais questões quando tais tipos de história lhe são apresentadas. Quando se fornece uma história para a máquina e se formula uma questão, a máquina imprimirá respostas do mesmo tipo que esperaríamos de seres humanos. Partidários da IA no sentido forte afirmam desta sequência pergunta-resposta que não somente a máquina está simulando uma habilidade humana, mas também que:

1. A máquina *compreende* a história e fornece respostas às questões,

2. O que a máquina e seu programa fazem *explica* a habilidade humana para entender histórias e responder questões sobre elas.

Ambas as afirmações me parecem totalmente insustentáveis a partir do trabalho de Schank[1], como tentarei mostrar no que se segue.

Uma maneira para testar qualquer teoria da mente é perguntar a alguém o que aconteceria se sua própria mente de fato funcionasse sob os princípios que a teoria diz que toda mente funciona. Vamos aplicar este teste ao programa de Schank com o seguinte

[1] Não estou dizendo, é claro, que o próprio Schank está comprometido com essas afirmações.

Gedankenexperiment[♣]. Suponha que estou trancado em um quarto e suponha que me dão um calhamaço de papel com um texto em chinês. Além disso, suponha que eu não conheça o idioma chinês, nem escrito nem falado, e que eu não seja sequer capaz de reconhecer a escrita chinesa, ou seja, distingui-la, por exemplo, da escrita japonesa ou de rabiscos sem significado. Suponha, agora, que além deste primeiro calhamaço fornecem-me – também em chinês – um segundo, contendo um roteiro com um conjunto de regras para correlacionar o segundo texto com o primeiro. As regras são em inglês e eu as compreendo tão bem como qualquer outro falante nativo de inglês. Isto me possibilita relacionar um conjunto de símbolos formais com o outro, e o que entendo por "formal" aqui é que posso identificar os símbolos por seu formato. Nestas circunstâncias imagine também que me forneçam um terceiro calhamaço contendo símbolos em chinês junto com algumas instruções, outra vez em inglês, as quais me possibilitarão correlacionar elementos deste terceiro maço com os dois primeiros; estas regras me instruem a como relacionar determinados símbolos em chinês com certos tipos de configuração e os devolver como resposta a determinadas configurações dadas no terceiro calhamaço. Sem que eu saiba, as pessoas que me fornecem os textos com os referidos símbolos denominam o primeiro bloco de "roteiro", o segundo, de "história" e o terceiro, de "questões". Ademais, eles intitulam os símbolos devolvidos em resposta ao terceiro maço de "respostas às questões", e o conjunto de regras em inglês, de "programa". Para complicar a história um pouquinho mais, imagine que estas pessoas também me forneçam histórias em inglês, as quais eu compreendo, e então elas me fazem questões em inglês sobre estas histórias, e eu as devolvo respondendo em inglês. Suponha, ainda, que depois de um tempo eu me saia tão bem ao seguir as instruções para manipulação dos símbolos em chinês e que os programadores consigam escrever tão bem os programas que do ponto de vista externo – isto é, do ponto de vista de alguém que esteja do lado de fora do quarto no qual eu estou trancado – minhas respostas às questões são indistinguíveis daquelas de falantes nativos de chinês. Ninguém observando minhas respostas pode dizer que eu não falo uma palavra de chinês. Vamos também supor que minhas respostas às questões em inglês sejam indistinguíveis daquelas de outro falante nativo de inglês – pela simples razão de que eu sou um falante nativo de inglês. Do ponto de vista externo – na visão de alguém que lê minhas "respostas" –, as respostas em chinês e em inglês são igualmente satisfatórias. Mas, no caso do idioma chinês, eu obtenho respostas manipulando símbolos

[♣] N. T.: O termo alemão *Gedankenexperiment* significa "experimento mental", um recurso filosófico onde se imagina uma situação possível que não contraria possibilidades físicas e lógicas e da qual podemos extrair consequências conceituais importantes.

formais em chinês, sem significação. No que diz respeito ao chinês, eu simplesmente me comportei como um computador; executei operações computacionais com base em elementos formalmente especificados. Para os propósitos do idioma chinês, eu sou simplesmente uma instanciação de um programa de computador.

Assim sendo, as afirmações feitas pela IA no sentido forte são de que um computador programado entende as histórias e que o programa, em algum sentido, explica a compreensão humana. Estamos agora em posição de examinar claramente estas afirmações no nosso experimento mental.

1. Considerando a primeira afirmação, parece óbvio no exemplo acima que eu não compreendo uma palavra das histórias em chinês. Eu tenho inputs e outputs que são indistinguíveis para os falantes nativos de chinês e, mesmo que eu tenha qualquer programa formal, ainda assim eu não compreendo nada. Pelas mesmas razões, o computador de Schank não compreende nada das histórias, sejam elas em chinês, em inglês, ou em qualquer outro idioma. No caso do idioma chinês, eu desempenho o papel do computador, e nos casos onde não desempenho tal papel, o computador não faz nada além do que eu poderia fazer, ou seja, em ambas as situações não há compreensão.

2. Com relação à segunda afirmação – que o programa explica a compreensão humana – podemos verificar que o computador e seu programa não fornecem as condições suficientes para a compreensão, visto que o computador e o programa estão funcionando e não existe compreensão. Mas será que ele fornece uma condição necessária ou uma contribuição significativa para a compreensão? Uma das afirmações sustentada pela IA no sentido forte é esta: quando eu compreendo uma história em inglês, o que estou fazendo é exatamente o mesmo – ou talvez mais que o mesmo – que fazia no caso da manipulação dos símbolos em chinês. No caso do inglês, que eu compreendo, há muito mais do que manipulação de símbolos formais do que em relação ao chinês, que eu não compreendo. Não estou demonstrando que esta afirmação é falsa, mas certamente me parece sem credibilidade no exemplo. A plausibilidade de tal suposição deriva do fato de que podemos construir um programa que terá os mesmos inputs e outputs como um falante nativo, além disso pressupomos que falantes têm algum nível de descrição onde eles são também instanciações de um programa. Com base nestas duas suposições, assumimos que, mesmo se o programa de Schank não constituir uma explicação completa da compreensão, talvez constitua uma parte de tal explicação. Ou seja, assumimos como possibilidade empírica, embora sem razões para supor que ela seja verdadeira (uma vez que ela é apenas sugerida e não demonstrada), que o programa de computador é irrelevante para

minha compreensão da história. No caso do idioma chinês, tenho tudo que a IA poderia colocar em mim por intermédio de um programa, e mesmo assim não compreendo nada. No caso do inglês, compreendo tudo e até agora não tenho nenhuma razão para supor que minha compreensão tenha alguma relação com programas de computador – isto é, com operações computacionais especificadas sobre elementos puramente formais. Na medida em que o programa é definido em termos de operações computacionais baseadas em elementos puramente formais, o que o exemplo sugere é que estes não têm conexão com a compreensão. Eles não são condição suficiente e não há, tampouco, razão para supor que eles sejam condição necessária ou mesmo que eles tenham alguma contribuição significativa para a compreensão. Observe-se que a força do argumento não é simplesmente que máquinas diferentes podem ter o mesmo input e output enquanto operando em princípios formais diferentes – não é este o ponto. O que queremos dizer é que, por mais que se coloque no computador princípios formais, isto não será suficiente para a compreensão, uma vez que um ser humano será capaz de seguir tais princípios formais sem compreender nada. Não há vantagem em supor que eles sejam necessários ou mesmo que contribuam em algo, visto que não há nenhuma razão para supor que, quando eu compreendo inglês, estou operando com algum programa formal.

O que há no caso das sentenças em inglês que não existe no caso das sentenças em chinês? A resposta óbvia é que eu sei o que as primeiras significam, mas não tenho a menor ideia do que as últimas significam. No que isto consiste e por que não posso atribuí-lo a uma máquina, qualquer que seja ela? Por que não posso atribuir a uma máquina aquilo que faz com que eu saiba o que as sentenças em inglês significam? Voltarei a estas questões depois de desenvolver um pouco mais o meu exemplo.

Tive a oportunidade de apresentar este exemplo a vários pesquisadores da IA e, curiosamente, eles parecem discordar acerca do que seja uma resposta para estas questões. Obtive uma variedade surpreendente de respostas, e, no que se segue, analisarei várias delas (especificadas conforme suas origens geográficas).

Primeiro, entretanto, quero desmontar alguns equívocos comuns sobre “compreensão”. Em muitas destas discussões, encontramos muita confusão sobre a palavra “compreensão”. Meus críticos alegam que há diferentes graus de compreensão, que “compreensão” não é um simples predicado binário, que existem de fato diferentes tipos e níveis de compreensão e, frequentemente, a lei do terceiro excluído não se aplica de uma maneira direta a enunciados da forma “x compreende y”; em muitos casos, se x compreende y é matéria de decisão e não uma simples questão de fato, e assim por diante. Sobre todos estes comentários, eu digo: “está certo,

é isso mesmo", mas eles não têm nada a ver com o que está sendo discutido aqui. Há casos em que "compreensão" se aplica claramente e casos onde claramente ela não se aplica. São situações deste tipo que preciso para fundamentar meu argumento[2]. Compreendo histórias em inglês; em grau inferior, posso também compreender histórias em francês; em um grau ainda menor, alemão; e em chinês, de jeito nenhum. Meu carro e minha máquina de somar, por um outro lado, não compreendem nada, estão "por fora" seja por metáfora ou por analogia. Frequentemente atribuímos "compreensão" e outros predicados cognitivos a carros, máquinas de somar e outros artefatos, mas nada se prova com tais atribuições. Dizemos: "a porta *sabe* quando abrir, em razão de sua célula fotoelétrica"; "a máquina de somar *sabe como* fazer soma e subtração, mas não divisão" e "o termostato *percebe* as mudanças de temperatura". A razão pela qual fazemos estas atribuições é interessante e tem a ver com o fato de que estendemos nossa própria intencionalidade para os artefatos[3]. Nossos instrumentos são extensões de nossos propósitos, e assim achamos natural fazer atribuições metafóricas de intencionalidade a eles; mas estes exemplos não resolvem nosso problema filosófico. O sentido no qual uma porta automática "compreende instruções" através de sua célula fotoelétrica não é de jeito nenhum o sentido no qual eu compreendo inglês. Se o sentido da compreensão de histórias dos computadores programados por Schank fosse o sentido metafórico no qual a porta compreende e não o sentido no qual eu compreendo inglês, não valeria a pena discutir este problema. Newell e Simon (1963) escrevem afirmando que o sentido de "compreensão" para os computadores é exatamente o mesmo que para os seres humanos. Gosto do modo incisivo desta afirmação e é este tipo de asserção que analisarei. Argumentarei que, em um sentido literal, o computador não compreende nada, da mesma maneira que o carro e a máquina de somar também não compreendem nada. A compreensão do computador não é como minha compreensão de alemão, ou seja, parcial ou incompleta, ela é zero.

Examinemos agora as objeções:

I. A objeção dos sistemas (Berkeley). "Embora seja verdade que a pessoa que está trancada no quarto não compreende a história, ocorre que ela é meramente parte de um sistema

[2] "Compreensão" implica não só a posse de estados mentais (intencionais) como também as condições de verdade desses estados (validade, sucesso). No escopo desta discussão, estamos interessados somente na posse desses estados.

[3] Intencionalidade é por definição aquela característica de determinados estados mentais pelos quais eles são direcionados para, ou acerca de objetos e estados de coisas no mundo. Neste caso, crenças, desejos e intenções são estados intencionais; formas não direcionadas de ansiedade e de depressão não são. Para uma discussão adicional ver Searle (1979).

global, e o sistema compreende a história. Essa pessoa tem uma grande tabela à sua frente na qual estão escritas as regras, tem um bloco de papel de rascunho, lápis para fazer cálculos; além disso, tem um 'banco de dados' com um conjunto de símbolos em chinês. Assim sendo, a compreensão não deve ser atribuída a um simples indivíduo, mas à totalidade de um sistema do qual ele faz parte".

Minha resposta à teoria dos sistemas é simples: deixe o indivíduo internalizar todos estes elementos do sistema. Ele memoriza as regras da tabela e o banco de dados com símbolos chineses, e então ele fará todos os cálculos em sua cabeça. O indivíduo, desse modo, incorpora todo o sistema. Não há nada no sistema que ele não possa abarcar. Podemos até dispensar o quarto e supor que ele trabalha do lado de fora. Do mesmo jeito, ele continuará não compreendendo nada de chinês; portanto, o sistema não compreende nada porque não há nada neste sistema que não esteja nele. Se ele não compreende, então o sistema não poderá compreender, pois o sistema é somente uma parte dele.

Na realidade, sinto-me até embaraçado ao dar uma resposta à teoria dos sistemas. A ideia é que, embora uma pessoa não compreenda chinês, de alguma forma a *conjunção* pessoa e pedacinhos de papel poderia compreender chinês. Não é fácil para mim imaginar como alguém que não estivesse preso a uma ideologia acharia esta ideia plausível. Entretanto, penso que muita gente que está comprometida com a ideologia da IA no sentido forte estará propensa a dizer algo muito parecido com isto. Vamos então explorar um pouco mais esta ideia. De acordo com uma versão desta visão, enquanto o homem do exemplo dos sistemas internalizados não compreende chinês como um falante nativo o faz (pois, por exemplo, ele não sabe que a história se refere a restaurante e hambúrgueres etc.), ainda assim "o homem como sistema de manipulação de símbolos formais" *realmente compreende chinês*. O subsistema do homem, que é o sistema de manipulação de símbolos formais para o chinês, não deve ser confundido com o subsistema para inglês.

Assim sendo, existem dois subsistemas no homem, um compreende inglês, o outro, chinês, e "acontece que os dois sistemas têm muito pouco a ver um com o outro". Mas quero responder que não somente eles têm muito pouco a ver um com o outro, como eles não são nem remotamente parecidos. O subsistema que compreende inglês (supondo que possamos usar este jargão "subsistema" no momento) sabe que as histórias são sobre restaurantes e comer hambúrgueres etc., ele sabe que estão formulando questões sobre restaurantes e que ele as responde da melhor maneira possível através de várias inferências sobre o conteúdo da história, e assim por diante. Mas o sistema chinês não sabe nada disso; enquanto o subsistema inglês

sabe que “hambúrgueres” se referem a hambúrgueres, o sistema chinês sabe somente que “tal e tal rabisco” é seguido de “outro rabisco”. Tudo o que ele sabe é que vários símbolos formais estão sendo introduzidos numa das extremidades e são manipulados de acordo com regras escritas em inglês, e que outros símbolos estão saindo na outra extremidade. O ponto essencial do exemplo original era argumentar que tal manipulação de símbolos por si só não poderia ser suficiente para compreender chinês nem no sentido literal, porque o homem poderia escrever “tal e tal rabisco” e depois “outro rabisco tal e tal” sem entender nada de chinês. E não vem de encontro ao argumento postular subsistemas dentro do homem, pois tais subsistemas não se desempenham melhor do que o homem; eles não têm nem sequer alguma semelhança com o falante de inglês (ou subsistema). De fato, na descrição feita, o subsistema chinês é simplesmente uma parte do subsistema inglês, uma parte que processa uma manipulação de símbolos sem sentido de acordo com regras em inglês.

Perguntemo-nos em primeiro lugar o que motiva a objeção dos sistemas – ou seja, que fundamentos *independentes* existem para se dizer que o agente deve ter um subsistema dentro dele que literalmente compreende histórias em chinês? Pelo que sei, os únicos fundamentos são que no exemplo eu tenho o mesmo input e o mesmo output dos falantes nativos de chinês e um programa que os intermedeia. Mas o ponto do exemplo foi mostrar que isto não poderia ser suficiente para a compreensão no sentido no qual compreendo histórias em inglês, pois uma pessoa e, portanto, o conjunto de sistemas que a compõe pode ter a combinação adequada de input, output e programa e mesmo assim não compreender nada no sentido no qual compreendo inglês. A única motivação para dizer que *deve* haver um subsistema em mim que compreende chinês é que eu tenho um programa e que posso passar no teste de Turing; posso enganar falantes nativos de chinês. Mas precisamente um dos pontos em discussão é a adequação do teste de Turing. O exemplo mostra que pode haver dois “sistemas”, ambos passam no teste de Turing, mas apenas um deles compreende; e não é um argumento contra este ponto dizer que se ambos passam no teste de Turing, ambos devem compreender, uma vez que esta afirmação não vem ao encontro do argumento de que o sistema em mim que compreende inglês é muito mais completo do que o sistema que meramente processa chinês. Em suma, a objeção do sistema escamoteia a questão ao insistir em apresentar argumentos que o sistema deve compreender chinês.

Além do mais, a objeção dos sistemas parece levar a consequências absurdas. Se tenho de concluir que deve haver cognição em mim com base no fato de que tenho um certo tipo de input e de output e um programa entre estes, então parece que todos os tipos de subsistemas não

cognitivos tornar-se-ão cognitivos. Por exemplo, meu estômago tem um nível de descrição no qual faz processamento de informação e instancia um grande número de programas de computador, mas suponho que não queremos dizer que ele tem compreensão [cf. Pylyshyn: "Computation and Cognition" *BBS* 3(1), 1980]. Se aceitamos a objeção dos sistemas, fica difícil de perceber como poderíamos evitar de dizer que o estômago, o coração, o fígado etc. são todos subsistemas que compreendem, pois não haveria nenhuma maneira, em princípio, para distinguir a motivação para dizer que o subsistema chinês compreende de dizer que o estômago compreende. Não constitui uma resposta para este ponto dizer que o sistema chinês tem informação como input e output e que o estômago tem comida e produtos alimentares como input e output, pois, do ponto de vista do agente e do meu ponto de vista, não há informação nem na comida e nem no chinês; o chinês é só um conjunto de rabiscos sem significado. A informação no caso do chinês está somente nos olhos dos programadores e dos intérpretes, e não há nada que os impeça de tratar o input e o output de meus órgãos digestivos como informação, se eles assim o quiserem.

Este último ponto diz respeito a alguns problemas na IA no sentido forte e vale a pena fazer aqui uma pequena digressão. Se a IA no sentido forte é um ramo da psicologia, ela deve ser capaz de distinguir sistemas que são genuinamente mentais daqueles que não o são. Ela deve ser capaz de distinguir os princípios com os quais a mente trabalha daqueles com os quais sistemas não mentais trabalham; de outra maneira, ela não poderia oferecer explicações acerca da natureza do que é especificamente mental. A distinção mental e não mental não pode estar apenas no olho do observador – ela deve ser intrínseca aos sistemas, pois de outra maneira ficaria a critério do observador tratar pessoas como não mentais e furacões como mentais. Mas com muita frequência, na literatura sobre IA, a distinção é esmaecida de tal maneira que se torna desastroso afirmar que a IA é uma investigação cognitiva. McCarthy, por exemplo, escreve: "Podemos dizer que máquinas tão simples como os termostatos têm crenças, e ter crenças parece ser uma característica de muitas máquinas capazes de resolver problemas" (McCarthy, 1979). Qualquer um que pense que a IA no sentido forte tem alguma chance como uma teoria da mente deve ponderar as implicações desta observação. Pedem-nos para aceitar como sendo uma descoberta da IA no sentido forte que o pedaço de metal na parede que usamos para regular a temperatura tenha crenças da mesma maneira que nós, nossas esposas e nossos filhos têm crenças, e além do mais que a "maioria" das outras máquinas da sala – telefone, gravador, máquina de somar, interruptor elétrico etc. – também tenham crenças. Não é objetivo deste artigo argumentar ou discutir com McCarthy, então afirmaremos o seguinte, sem

argumentar. O estudo da mente começa com o fato de que seres humanos têm crenças e que termostatos, telefones e máquinas de somar não as têm. Se você concebe uma teoria que nega tal ponto, você produziu um contraexemplo e a teoria é falsa. Têm-se a impressão de que os pesquisadores da IA que escrevem esse tipo de coisa pensam que podem escapar disto porque eles realmente não levam tais coisas a sério e não pensam que alguém o fará. Proponho, pelo menos para o momento, levar estas coisas a sério. Pense por um minuto o que seria necessário para estabelecer que o pedaço de metal na parede tem, de fato, crenças – crenças com direcionalidade, conteúdo proposicional, condições de satisfação; crenças que têm a possibilidade de ser fortes ou fracas, ansiosas ou seguras, dogmáticas, racionais ou supersticiosas, fé cega ou especulações hesitantes. O termostato não é um candidato plausível a ter crenças, nem tampouco o são o estômago, o fígado, a máquina de somar ou o telefone. Contudo, uma vez que estamos levando esta ideia a sério, note-se que, se fosse verdadeira, ela seria fatal para a proposta da IA de ser uma ciência da mente, pois então a mente estaria em todos os lugares. O que queremos saber é o que distingue a mente de termostatos, fígados etc. Se McCarthy estivesse certo, a IA no sentido forte não teria a menor possibilidade de nos dizer em que se baseia esta distinção.

II. A objeção do robô – (Yale). "Suponhamos que escrevêssemos um programa diferente daquele de Schank. Suponhamos que puséssemos um computador dentro de um robô e que esse computador não fosse apenas receber símbolos formais como input e produzir esses símbolos como output, mas que ele fosse operar o robô de tal maneira que este fizesse coisas como perceber, andar, mover-se, pregar pregos, comer, beber ou qualquer outra coisa. O robô teria uma câmera de televisão adaptada a ele – o que o capacitaria a 'ver' – teria braços e pernas que o capacitariam a 'agir' e tudo isso seria controlado pelo seu cérebro-computador. Tal robô teria compreensão genuína e outros estados mentais – ele seria diferente do computador de Schank".

A primeira coisa a notar acerca da objeção do robô é que ela tacitamente concede que a cognição não é só uma questão de manipulação de símbolos, uma vez que esta objeção acrescenta um conjunto de relações causais com o mundo externo [cf. Fodor: "Methodological Solipsism" *BBS* 3(1), 1980]. Mas a resposta à objeção do robô é que o acréscimo de tais capacidades (perceptual e motora) não acrescenta nada em termos de compreensão ou intencionalidade ao programa original de Schank. Para perceber isso, basta notar que o mesmo experimento mental se aplica ao caso do robô. Suponha que, em vez de um computador dentro

de um robô, você me ponha dentro do quarto e me dê novamente símbolos em chinês com instruções em inglês para combinar estes símbolos com outros símbolos em chinês. Suponhamos que, sem eu saber, alguns dos símbolos em chinês que chegam a mim venham de uma câmera de televisão adaptada ao robô, e que outros símbolos em chinês que estou produzindo sirvam para fazer com que o motor dentro do robô mova seus braços e pernas. É importante enfatizar que tudo que estou fazendo é manipular símbolos formais. Estou recebendo "informação" do aparato "perceptual" do robô e estou fornecendo "instruções" para seu aparato motor sem saber o que estou fazendo. Eu sou o homúnculo do robô, mas, de maneira diferente do homúnculo tradicional, sem saber o que está ocorrendo. Não sei nada a não ser as regras para manipulação de símbolos. Neste caso, pode-se dizer que o robô não tem estados intencionais; ele se move como resultado de seus circuitos elétricos e do seu programa. Além do mais, a instanciação de um programa não produz estados intencionais de nenhum tipo relevante. Tudo o que está sendo feito é seguir instruções formais acerca da manipulação de símbolos formais.

III. A objeção do simulador cerebral (Berkeley e M.I.T.). "Suponhamos que nós projetássemos um programa que não represente a informação que temos acerca do mundo, como é o caso da informação dos roteiros de Schank. O programa simula a sequência efetiva da atividade dos neurônios nas sinapses do cérebro de um falante nativo de chinês quando este entende histórias e dá respostas a elas. A máquina recebe histórias em chinês e questões acerca delas como input; ela simula a estrutura formal dos cérebros dos chineses ao processar estas histórias e fornece respostas em chinês como outputs. Podemos até imaginar que a máquina não opera com um único programa serial, mas com um conjunto de programas operando em paralelo, da mesma maneira que cérebros humanos possivelmente operam quando processam linguagem natural. Em tal caso, teríamos de dizer que a máquina entenderia histórias; e se nos recusássemos a dizer isso, não teríamos também que negar que falantes de chinês entendem histórias? Ao nível das sinapses, o que poderá ser diferente no programa do computador e no programa do cérebro dos chineses?".

Antes de responder à esta objeção, quero fazer uma digressão para notar que esta é uma objeção estranha de ser feita por qualquer adepto da IA (funcionalismo etc.). Penso que a ideia central da IA no sentido forte é que não precisamos saber como o cérebro funciona para saber como a mente funciona. A hipótese básica é que existe um nível de operações mentais que consiste em processos computacionais sobre elementos formais que constitui a essência do

mental e pode ser realizado através de diferentes processos cerebrais, da mesma maneira que um programa computacional pode ser rodado em diferentes hardwares. A pressuposição da IA no sentido forte é que a mente está para o cérebro assim como o programa está para o hardware, e podemos entender a mente sem fazer neurofisiologia. Se tivéssemos que saber como o cérebro trabalha para fazer IA, esta não constituiria um problema. Contudo, mesmo que cheguemos a um conhecimento muito grande das operações do cérebro, isto não seria suficiente para produzir a compreensão. Senão, vejamos: imagine que ao invés de um ser monolingual num quarto combinando símbolos tenhamos um homem operando um conjunto complexo de canos de água com válvulas que os conectam. Quando o homem recebe símbolos em chinês, ele consulta no programa, escrito em inglês, quais válvulas ele deve abrir e quais ele deve fechar. Cada conexão na tubulação corresponde a uma sinapse no cérebro do chinês, e o sistema é equipado de tal maneira que após ativar as conexões adequadas – ou seja, após abrir as torneiras adequadas – as respostas em chinês apareçam no final da tubulação.

Onde está a compreensão neste sistema? Ele recebe chinês como input, simula a estrutura formal das sinapses do cérebro do chinês e produz textos em chinês como output. Mas o homem certamente não entende chinês, e nem tampouco a tubulação, e se estivermos tentados a adotar o que penso ser a ideia absurda de que de alguma maneira a *conjunção* homem *e* tubulação compreende, é preciso lembrar que em princípio o homem pode internalizar a estrutura formal da tubulação de água e realizar toda a "atividade neuronal" em sua imaginação. O problema com o simulador cerebral é que ele está simulando coisas erradas acerca do cérebro. Na medida em que ele simula unicamente a estrutura formal das sequências de atividades neuronais nas sinapses, ele não está simulando o aspecto mais importante do cérebro, ou seja, suas propriedades causais e sua habilidade para produzir estados intencionais. Que as propriedades formais não são suficientes para produzir propriedades causais é mostrado pelo exemplo da tubulação de água: podemos ter todas as propriedades formais sem que estas tenham sido derivadas das propriedades causais neurobiológicas relevantes.

IV. A objeção da combinação – (Berkeley e Stanford). "As três objeções anteriores podem não ser convincentes como uma refutação do contraexemplo do quarto chinês, mas se elas forem tomadas conjuntamente são convincentes e decisivas. Imagine um robô com um computador em forma de cérebro alojado em sua cavidade craniana; imagine que o computador está programado com todas as sinapses de um cérebro humano; imagine que o comportamento do robô é indistinguível do comportamento humano e agora pense nisto tudo como um sistema

unificado e não apenas como um computador com inputs e outputs. Certamente em tal caso teríamos que atribuir intencionalidade ao sistema".

Concordo inteiramente que em tal caso acharíamos racional e mesmo irresistível aceitar a hipótese de que o robô teria intencionalidade, na medida em que não soubéssemos mais nada sobre ele. Além da aparência e comportamento, os outros elementos da combinação são irrelevantes. Se pudéssemos construir um robô cujo comportamento não se distinguisse de uma grande parcela do comportamento humano, nós lhe atribuiríamos intencionalidade, apesar de termos algumas razões para não o fazer. Não precisaríamos saber de antemão que seu cérebro-computador é um análogo formal do cérebro humano.

Mas realmente não vejo como isto poderia ajudar nas pretensões da IA no sentido forte; e eis o porquê: de acordo com a IA no sentido forte, instanciar um programa formal com o input e o output adequados é condição suficiente e constitutiva da intencionalidade. Como Newell (1979) coloca, a essência do mental é a operação de um sistema de símbolos físicos. Mas as atribuições de intencionalidade que fazemos ao robô neste exemplo não têm nada a ver com programas formais. Elas são simplesmente baseadas na pressuposição de que, se o robô se parece e se comporta como nós, teríamos de supor – até prova em contrário – que ele deve ter estados mentais como os nossos que causam e se expressam no seu comportamento, bem como um mecanismo interno capaz de produzir tais estados mentais. Se soubéssemos como explicar seu comportamento independentemente, sem tais pressuposições, não atribuiríamos intencionalidade a ele, especialmente se soubéssemos que ele tem um programa formal. Este é o ponto de minha resposta à objeção II.

Suponhamos que nós soubéssemos que o comportamento do robô é inteiramente explicado pelo fato de um homem dentro dele estar recebendo símbolos formais sem interpretação dos receptores sensoriais do robô e enviando esses símbolos para os mecanismos motores desse robô, e que o homem está fazendo essa manipulação simbólica de acordo com um conjunto de regras. Além do mais, suponha que o homem nada sabe desses fatos acerca do robô; tudo o que ele sabe é qual operação realizar sobre esses símbolos sem significado. Em tal caso, consideraríamos o robô como um engenhoso fantoche mecânico. A hipótese de que o fantoche tenha uma mente seria então injustificada e desnecessária, pois não haveria mais razão para atribuir intencionalidade ao robô ou para o sistema do qual ele é uma parte (com exceção da intencionalidade do homem que está manipulando os símbolos). A manipulação de símbolos formais continua, o input e o output são combinados corretamente, mas o único lócus de intencionalidade é o homem, e ele não sabe nada dos estados intencionais relevantes; por

exemplo, ele não *vê* o que chega aos olhos do robô, ele não tem a *intenção* de mover o braço do robô, ele não *compreende* as observações que são feitas pelo robô ou que lhe são feitas. Nem tampouco, pelas razões colocadas acima, o sistema do qual o homem e o robô são parte compreende alguma coisa.

Para esclarecer este ponto, façamos um contraste com os casos onde achamos completamente natural atribuir intencionalidade a membros de algumas outras espécies, como a gorilas e macacos e a animais domésticos como os cães. As razões pelas quais achamos isto natural são, *grosso modo*, duas. Sem atribuir intencionalidade aos animais, seu comportamento não faz sentido, e podemos ver que os animais são feitos de material semelhante ao nosso: olhos, nariz, pele etc. Dada a coerência do comportamento animal e a pressuposição de um mesmo material causal subjacente a ele, pressupomos que o animal deve ter estados mentais subjacentes a seu comportamento e que esses estados mentais devem ser produzidos por mecanismos feitos de um material semelhante ao nosso. Certamente poderíamos fazer pressuposições semelhantes acerca do robô, mas, na medida em que soubéssemos que seu comportamento resulta de um programa formal e que as propriedades causais efetivas de sua substância física fossem irrelevantes, abandonaríamos a pressuposição de intencionalidade [ver "Cognition and Consciousness in Nonhuman Species", *BBS* I(4), 1978].

Existem outras duas respostas ao meu exemplo que aparecem frequentemente (e então valeria a pena discuti-las), mas elas realmente fogem ao ponto.

V. A objeção das outras mentes (Yale). "Como saber que outras pessoas compreendem chinês ou qualquer outra coisa? Unicamente por seus comportamentos. Ora, o computador pode passar por testes de comportamento tão bem quanto elas (em princípio), assim se atribuímos cognição a outras pessoas, devemos em princípio atribuí-la também a computadores".

A objeção merece apenas uma resposta curta. O problema em questão não é como eu sei que outras pessoas têm estados cognitivos, mas o que estou lhes atribuindo ao dizer que elas têm estados cognitivos. O ponto central do argumento é que não poderiam ser apenas processos computacionais e seus outputs porque estes podem existir sem o estado cognitivo. Não é resposta para este argumento fingir que estados cognitivos não existem. Em "ciências cognitivas", pressupõe-se a realidade e a possibilidade de se conhecer o mental, da mesma maneira que em ciências físicas temos de pressupor a realidade e a capacidade de se conhecer objetos físicos.

VI. A objeção das "várias casas" (Berkeley). "A totalidade de seu argumento pressupõe que a IA trata apenas de computadores analógicos e digitais. Ocorre que este é apenas o estágio atual da tecnologia. Quaisquer que sejam esses processos causais que você diz serem essenciais para a intencionalidade (pressupondo que você esteja correto), possivelmente seremos capazes de construir dispositivos que exibirão esses processos causais e isto será também inteligência artificial. Assim, seus argumentos não se aplicam à capacidade da IA para produzir e explicar a cognição".

Não tenho resposta a esta objeção a não ser dizer que ela trivializa o projeto da IA no sentido forte ao redefini-la como qualquer coisa que produza e explique a cognição artificialmente. O interesse das afirmações originais feitas em favor da IA é que ela era uma tese precisa e bem definida: processos mentais são processos computacionais sobre elementos formalmente definidos. Minha preocupação tem sido desafiar esta tese. Se sua proposta é redefinida de tal maneira que ela não mais se constitui nesta tese, minhas objeções não se aplicam mais, pois não há mais uma hipótese testável sobre a qual elas se aplicam.

Retornemos às questões as quais prometi que tentaria responder. Dado que no exemplo original eu compreendo inglês e não chinês, e dado que a máquina não compreende nem inglês nem chinês, deve haver algo em mim que faz com que eu compreenda inglês e algo que falta em mim que faz com que eu não compreenda chinês. Por que não podemos dar essas coisas, sejam lá o que forem, a uma máquina?

Não vejo razão, em princípio, porque não poderíamos conceder a uma máquina a capacidade de compreender inglês ou chinês, pois nossos corpos com nossos cérebros são precisamente tais máquinas. Não há argumentos fortes para dizer que não poderíamos atribuir tal coisa a uma máquina se sua operação for definida somente em termos de processos computacionais sobre elementos formalmente definidos, ou seja, onde a operação da máquina é definida como uma instanciação de um programa de computador. Não é porque eu sou a instanciação de um programa de computador que eu sou capaz de entender inglês e ter outras formas de intencionalidade (eu sou, suponho, a instanciação de qualquer programa de computador), mas, pelo que sabemos, é porque eu sou um certo tipo de organismo com uma certa estrutura biológica (i.e., física e química), e esta estrutura, em termos causais, é capaz, sob certas condições, de produzir a percepção, a ação, a compreensão, o aprendizado e outros fenômenos intencionais. Parte do núcleo deste argumento é que só algo que tenha estes poderes causais pode ter intencionalidade. Talvez outros processos físicos e químicos pudessem

produzir exatamente estes efeitos; talvez, por exemplo, os marcianos também tenham intencionalidade, mas os seus cérebros sejam feitos de um material diferente. Esta é uma questão empírica, semelhante à questão de se a fotossíntese pode ser feita com uma química diferente da que compõe a clorofila.

Mas o ponto principal do presente argumento é que um modelo puramente formal nunca será, por si só, suficiente para produzir intencionalidade, pois as propriedades formais não são constitutivas da intencionalidade e não têm poderes causais, com exceção do poder de produzir o estágio seguinte do formalismo quando a máquina está rodando. E mesmo que uma realização específica do modelo formal venha a exibir propriedades causais, estas são irrelevantes, pois este modelo pode também ser efetivado através de uma realização diferente onde tais propriedades estarão ausentes. Mesmo que, por algum milagre, falantes de chinês realizem exatamente o programa de Schank, podemos colocar o mesmo programa em falantes de inglês, na tubulação de água ou em computadores; nenhum destes compreende chinês e nem tampouco o programa.

O que importa nas operações do cérebro não é a sombra do formalismo dado pela sequência das sinapses, mas as propriedades efetivas de tais sequências. Todos os argumentos em favor da versão forte da IA que examinei insistem em delinear estas sombras lançadas pela cognição para então sustentar que tais sombras são a própria cognição.

Com o intuito de concluir, quero enunciar alguns pontos filosóficos gerais implícitos no argumento. Por uma questão de clareza, tentarei fazer isto na forma de perguntas e respostas e começo com a velha questão:

"Pode uma máquina pensar?".

A resposta é, obviamente, sim. Nós somos precisamente tais máquinas.

"Sim, mas pode um artefato, uma máquina feita pelo homem, pensar?".

Assumindo que seja possível produzir artificialmente uma máquina com sistema nervoso, neurônios com axônios e dendritos e tudo o mais, suficientemente semelhante a nós, de novo a resposta a esta questão parece ser, obviamente, "sim". Se você pode duplicar exatamente as causas, você pode duplicar os efeitos. E de fato seria possível produzir consciência, intencionalidade e tudo o mais usando princípios químicos diferentes dos usados por seres humanos. Como eu disse, é uma questão empírica.

"O.K., mas pode um computador digital pensar?".

Se por um "computador digital" queremos dizer algo que tem um nível de descrição através do qual esse algo pode corretamente ser descrito como a instanciação de um programa

de computador, então de novo a resposta é sim, uma vez que somos as instanciações de um grande número de programas de computador e podemos pensar.

"Mas pode algo pensar, compreender etc. *somente* em virtude de ser um computador com o tipo de programa adequado? Pode a instanciação de um programa, de um programa adequado é claro, ser por si só condição suficiente para compreensão?".

Esta para mim é a questão correta a ser formulada, embora seja usualmente confundida com uma ou mais das questões anteriores, e a resposta para ela é "não".

"Por que não?".

Porque a manipulação de símbolos formais por si só não tem intencionalidade: eles não têm significado, eles nem mesmo são manipulações de *símbolos*, uma vez que esses símbolos não simbolizam nada. No jargão linguístico, eles têm apenas sintaxe, mas não semântica. A intencionalidade que os computadores parecem ter está apenas nas mentes daqueles que os programam e daqueles que os usam, ou seja, de quem envia o input e interpreta o output.

O objeto do exemplo do quarto chinês foi tentar mostrar isso, pois, na medida em que colocamos algo no sistema que realmente tem intencionalidade (um ser humano) e o programamos com o programa formal, pode-se ver que este programa não exibe intencionalidade adicional. Por exemplo, isto nada acrescenta à habilidade do ser humano para compreender chinês.

Precisamente, a característica da IA que parece tão atrativa – a distinção entre programa e realização – mostra-se fatal para a proposta de que simulação possa ser duplicação. A distinção entre o programa e sua realização no hardware encontra paralelo na distinção entre o nível de operações mentais e o nível de operações cerebrais. E se pudéssemos descrever o nível de operações mentais como um programa formal, poderíamos descrever o que é essencial acerca da mente sem fazer psicologia introspectiva ou neurofisiologia do cérebro. Mas a equação: "a mente está para o cérebro assim como o programa está para o hardware" tropeça em vários pontos, entre eles, os três seguintes:

Primeiro, a distinção entre programa e realização tem a consequência de que o mesmo programa poderia ter vários tipos de realizações absurdas sem nenhuma forma de intencionalidade. Weizenbaum (1976, cap. 2), por exemplo, mostra em detalhes como construir um computador usando um rolo de papel higiênico e uma pilha de pedrinhas. Similarmente, o programa para compreensão de histórias em chinês pode ser programado numa sequência de canos de água, num conjunto de cata-ventos ou num falante monolingual de inglês, nenhum dos quais, entretanto, adquire uma compreensão de chinês. Pedras, papel higiênico, vento e

canos de água são o material errado para gerar intencionalidade (apenas algo que tenha os mesmos poderes causais do cérebro pode ter intencionalidade), e embora o falante de inglês tenha o material correto para a intencionalidade, pode-se ver facilmente que ele não adquire nenhuma intencionalidade extra por memorizar o programa, uma vez que memorizá-lo não vai lhe ensinar chinês.

Segundo, o programa é puramente formal, mas os estados intencionais não são formais. São definidos em termos de seu conteúdo, e não de sua forma. A crença de que está chovendo, por exemplo, não é definida como uma determinada configuração formal, mas como um determinado conteúdo mental com condições de satisfação, de racionalidade etc. (ver Searle, 1979). Com efeito, a crença como tal não tem sequer uma configuração formal no sentido sintático, uma vez que a apenas uma e mesma crença pode ser dado um número indefinido de expressões sintáticas diferentes em diferentes sistemas linguísticos.

Terceiro, como mencionei anteriormente, estados e eventos mentais são produtos da operação do cérebro, mas o programa não é um produto do computador.

"Bem, se os programas não são constitutivos de processos mentais, por que tantas pessoas acreditaram no oposto? Isso precisa ser explicado".

Não sei a resposta para isto. A ideia de que as simulações computacionais poderiam ser a própria mente deve ter parecido suspeita, em primeiro lugar, porque o computador de nenhuma maneira se limita a simular operações mentais. Ninguém supõe que simulações computacionais de um alarme contra fogo causarão um incêndio na vizinhança ou que uma simulação computacional de uma tempestade deixar-nos-á encharcados. Por que alguém suporia então que uma simulação computacional da compreensão de fato entenderia alguma coisa? Diz-se frequentemente que seria extremamente difícil fazer computadores sentirem dor ou se apaixonarem, mas amor e dor não são nem mais fáceis nem mais difíceis de simular do que a cognição ou qualquer outra coisa. Para fazer uma simulação, tudo que se precisa é um input e um output corretos e um programa que os intermedeie, transformando o primeiro no segundo. Isto é tudo o que o computador tem e tudo o que ele pode fazer. Confundir simulação com duplicação é o mesmo erro, seja com dor, amor, cognição, incêndio ou tempestade.

Mesmo assim, há várias razões pelas quais a IA deve ter parecido – e para muitas pessoas ainda parece – reproduzir e explicar fenômenos mentais, e acredito que não conseguiremos remover estas ilusões até que tenhamos exposto as razões que as originaram.

Em primeiro lugar, e talvez o mais importante, está a confusão a respeito da noção de "processamento de informação". Muitas pessoas, em ciência cognitiva, acreditam que o cérebro

humano, com sua mente, faz algo chamado “processamento de informação”, e analogamente o computador com seu programa faz processamento de informação; mas, por outro lado, incêndios e tempestades não o fazem. Embora o computador possa simular aspectos formais de qualquer tipo de processo, ele está numa relação especial com a mente e o cérebro, pois quando o computador é adequadamente programado, idealmente com o mesmo programa do cérebro, o processamento de informação é idêntico nos dois casos, e este processamento de informação é realmente a essência do mental. Mas o problema com este argumento é que ele repousa sobre uma ambiguidade na noção de “informação”. O sentido pelo qual as pessoas “processam informação” quando elas refletem sobre problemas aritméticos, ou quando elas leem e respondem questões sobre histórias, não é o sentido no qual o computador programado “processa informação”. Em vez disso, o que ele faz é manipular símbolos formais. O fato de que o programador e o intérprete dos outputs do computador usem símbolos para representar objetos do mundo está totalmente além do escopo do computador. Repetindo, o computador tem sintaxe, mas não tem semântica. Dessa forma, se você digita: “2+2 igual?”, ele vai apresentar “4”. Mas ele não tem ideia de que “4” significa 4, ou que isto signifique alguma coisa. O ponto não é que ele não tenha alguma informação de segunda ordem acerca da interpretação de seus símbolos de primeira ordem, mas o fato é que estes símbolos de primeira ordem não têm nenhuma interpretação no que diz respeito ao computador. Tudo que ele tem são mais símbolos. Assim sendo, a introdução da noção de “processamento de informação” produz um dilema: ou bem construímos a noção de “processamento de informação” de tal maneira que ela implique a intencionalidade como parte do processo, ou bem nós não o fazemos. No primeiro caso, então, o computador programado não processa informação, ele somente manipula símbolos formais. No segundo caso, então, apesar de o computador processar informação, é somente no sentido em que máquinas de somar, máquinas de escrever, estômagos, termostatos, tempestades e furacões o fazem – a saber, eles têm um nível de descrição no qual podemos descrevê-los como recebendo informação, transformando-a e produzindo informação como output. Mas nesse caso depende de observadores externos interpretar o input e o output como informação no sentido comum. E nenhuma semelhança é estabelecida entre o computador e o cérebro em termos de uma similaridade de processamento de informação nos dois casos.

Em segundo lugar, em grande parte da IA há um behaviorismo residual ou operacionalismo. Uma vez que computadores adequadamente programados podem ter padrões de input-output semelhantes ao de seres humanos, somos tentados a postular estados mentais

no computador similares a estados mentais humanos. Mas, uma vez que percebemos que é conceitual e empiricamente possível para um sistema ter capacidades humanas em algum domínio sem ter nenhuma intencionalidade, devemos ser capazes de superar este impulso. Minha máquina de somar tem capacidade de calcular, mas não intencionalidade, e neste artigo tentei mostrar que um sistema pode ter capacidades de input e output que duplicam aquelas de um falante nativo de chinês e ainda assim não compreender chinês, a despeito de como ele é programado. O teste de Turing é típico na tradição de ser abertamente behaviorista e operacionalista, e acredito que se os pesquisadores da IA repudiassem totalmente o behaviorismo e o operacionalismo, muito da confusão entre simulação e duplicação seria eliminada.

Em terceiro lugar, este operacionalismo residual junta-se a uma forma residual de dualismo; de fato, a IA no sentido forte só faz sentido com uma pressuposição dualista onde aquilo que diz respeito à mente nada tem a ver com o cérebro. Na IA no sentido forte (bem como no funcionalismo), o que importa são programas, e programas são independentes de sua realização em máquinas; de fato, no que diz respeito à IA, um mesmo programa pode ser realizado por uma máquina eletrônica, uma substância mental cartesiana ou o espírito do mundo hegeliano. A descoberta mais surpreendente que eu fiz ao discutir estes problemas é que muitos pesquisadores da IA estão chocados com a minha ideia de que fenômenos mentais humanos podem ser dependentes das efetivas propriedades físico-químicas dos cérebros humanos. Mas eu não deveria estar surpreso, pois, a não ser que se aceite alguma forma de dualismo, o projeto da IA no sentido forte não tem nenhuma chance. O projeto consiste em reproduzir e explicar o mental projetando programas, mas a não ser que a mente seja não apenas conceitual, mas também empiricamente independente do cérebro, este projeto não poderá ser executado, pois o programa é completamente independente de qualquer realização. A não ser que se acredite que a mente é separável do cérebro conceitual e empiricamente – um dualismo em uma versão forte –, não se pode esperar reproduzir o mental escrevendo e rodando programas, uma vez que estes devem ser independentes dos cérebros ou de qualquer outra forma específica de sua instanciação. Se operações mentais consistem em operações computacionais sobre símbolos formais, segue-se que eles não têm nenhuma conexão importante com o cérebro, e a única conexão seria que o cérebro poderia ser um dentre os múltiplos tipos de máquinas capazes de instanciar o programa. Esta forma de dualismo não é a versão cartesiana tradicional, a qual sustenta que existem dois tipos de *substâncias*, mas é cartesiana no sentido de que ela insiste que aquilo que é especificamente mental não tem nenhuma conexão intrínseca com as

propriedades efetivas do cérebro. Este dualismo subjacente é mascarado pelo fato de que a literatura sobre IA contém frequentes ataques contra o "dualismo", mas o que estes autores não percebem é que sua posição pressupõe uma versão forte do dualismo.

"Pode uma máquina pensar?". Meu ponto de vista é que *somente* uma máquina pode pensar, e de fato apenas máquinas de um tipo muito especial, a saber, cérebros e máquinas que têm os mesmos poderes causais do cérebro. E esta é a principal razão pela qual a IA no sentido forte tem tão pouco a dizer acerca do pensamento: ela não tem nada a dizer acerca de máquinas. Por definição, ela trata de programas, e programas não são máquinas. O que quer que seja a intencionalidade, é um fenômeno biológico, o qual deve ser tão causalmente dependente da bioquímica específica de suas origens como o é a lactação, a fotossíntese ou quaisquer outros fenômenos biológicos. Ninguém suporia que poderíamos produzir leite e açúcar rodando uma simulação computacional das sequências formais da lactação e da fotossíntese; mas no que diz respeito à mente muitas pessoas querem acreditar em tal milagre por causa de sua fidelidade profunda ao dualismo: concebem a mente como processos formais e como algo independente de causas materiais específicas, algo que não ocorre com o açúcar e o leite.

Na defesa desse dualismo, expressam essa esperança na forma de que o cérebro é um computador digital (computadores antigos eram frequentemente chamados de "cérebros eletrônicos"). Mas isto não adianta nada. É claro que o cérebro é um computador digital. Uma vez que tudo é um computador digital, os cérebros também o são. O ponto é que a capacidade causal do cérebro para produzir intencionalidade não pode consistir na instanciação de um programa de computador, pois, para qualquer programa, sempre é possível que haja algo que o instancie e, contudo, não tenha estados mentais. Seja lá o que o cérebro faça para produzir intencionalidade, esta não pode consistir na instanciação de um programa, pois nenhum programa é por si só suficiente para produzir a intencionalidade.

Agradecimentos

Estou em débito com um grande número de pessoas que discutiram este assunto e por seu paciente esforço em superar minha ignorância em IA. Gostaria de agradecer especialmente a Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky e Terry Winograd.

Referências[♠]

FODOR, J. A. Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. *Behavioral and Brain Sciences* v. 3, n. 1, p. 63-73, March 1980. DOI: https://doi.org/10.1017/S0140525X00001771 Acesso em: 27 jan. 2021

McCARTHY, J. Ascribing Mental Qualities to Machines. *Stanford AI Lab Memo 326*, 1979. Disponível em: http://jmc.stanford.edu/articles/ascribing/ascribing.pdf Acesso em: 27 jan. 2021

NEWELL, A. & SIMON, H. A. GPS, a Program that Simulates Human Thought. In: FEIGENBAUM, A. & FELDMAN, V. (Eds.). *Computers and Thought*. New York: McGraw Hill, 1963, p. 279-93.

NEWELL, A. Physical Symbol Systems. *Lecture at the La Jolla Conference on Cognitive Science*, 1979. DOI: 10.1016/S0364-0213(80)80015-2 Acesso em: 27 jan. 2021 [NEWELL, A. Physical Symbol Systems. *Cognitive Science*, v. 4, n. 2, p. 135-83, April-June 1980, DOI: https://doi.org/10.1016/S0364-0213(80)80015-2 Acesso em: 27 jan. 2021].

PYLYSHYN, Z. W. Computation and Cognition: Issues in the Foundations of Cognitive Science. *Behavioral and Brain Sciences*, v. 3, n. 1, 111-69, 1980. DOI: 10.1017/S0140525X00002053 Acesso em: 27 jan. 2021

SCHANK, R. & ABELSON, R. P. Natural Language, Philosophy and Artificial Intelligence. In: RINGLE, M. (org.). *Philosophical Perspectives in Artificial Intelligence*. N. J.: Humanities Press, 1977.

SEARLE, J. What Is an Intentional State? *Mind*, v. 88, n. 1, p. 74-92, January 1979. DOI: 10.1093/mind/LXXXVIII.1.74 Acesso em: 27 jan. 2021

THE BEHAVIORAL AND BRAIN SCIENCES, v. 1, n. 4: *A Special Issue on Cognition and Consiousness in Nonhuman Species*, December 1978. Disponível em: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/issue/special-issue-on-cognition-and-consiousness-in-nonhuman-species/E7C03C97D42CE3CD7014BFD538402670 Acesso em: 27 jan. 2021

TURING, A. Computing Machinery and Intelligence. *Mind*, v. 59, n. 236, October 1950. DOI: https://doi.org/10.1093/mind/LIX.236.433

[♠] Nota dos Organizadores do Dossiê: acrescentamos as seguintes obras em relação à seção de Referências da Professora Cléa Regina de Oliveira Ribeiro: Fodor (1980), Newell & Simon (1963), Pylyshyn (1980), BBS [*The Behavioral and Brain Sciences*] (1978) e Winograd (1972). Com o intuito de auxiliar o leitor, sempre que possível, especificamos o DOI ou o *link* de acesso dos textos na presente republicação.

WEIZENBAUM, J. *Computer Power and Human Reason*. San Francisco: W.H. Freeman, 1976.

WINOGRAD, T. *Understanding Natural Language*. Massachusetts: Academic Press, INC., 1972.