MENTES, CÉREBROS E PROGRAMAS

John Searle

Que importância filosófica e psicológica se deve atribuir aos recentes esforços aplicados na simulação computacional das capacidades cognitivas humanas? Ao responder a esta questão, creio ser útil distinguir aquilo a que chamarei IA (inteligência artificial) «forte» de IA «fraca» ou «cautelosa». De acordo com a IA fraca, o principal valor do computador no estudo da mente é dar-nos uma poderosa ferramenta. Por exemplo, permite-nos formular e testar hipóteses de uma maneira mais precisa. Mas de acordo com a IA forte, o computador não é meramente uma ferramenta no estudo da mente; ao invés, o computador adequadamente programado é na verdade uma mente, no sentido em que se pode literalmente afirmar que, dados os programas adequados, os computadores *compreendem* e têm outros estados cognitivos. Na IA forte, porque o computador programado tem estados cognitivos, os programas não são meramente ferramentas que nos permitem testar explicações psicológicas; ao invés, os programas são em si explicações.

Nada tenho a objectar às afirmações da IA fraca, pelo menos no que diz respeito a este artigo. A discussão que faço aqui dirige-se às afirmações que defini como sendo as da IA forte, em especial a afirmação de que o computador adequadamente programado tem literalmente estados cognitivos e que, por isso, os programas explicam a cognição humana. Quando doravante me referir à IA, terei em mente a versão forte, como se exprime nestas duas afirmações.

Tomarei em consideração a obra de Roger Schank e dos seus colegas em Yale (Schank e Abelson, 1977), porque estou mais familiarizado com este do que com quaisquer outras afirmações similares e porque fornece um exemplo muito claro do tipo de trabalho que pretendo examinar. Mas nada do que se segue depende dos detalhes dos programas de Schank. Os mesmos argumentos aplicar-se-iam ao SHRDLU de Winograd (Winograd, 1973), ao ELIZA de Weizenbaum (Weizenbaum, 1965) e, com efeito, a qualquer simulação de fenómenos mentais humanos numa máquina de Turing.

Muito resumidamente, deixando de fora os diversos detalhes, pode-se descrever do seguinte modo o programa de Schank: o objectivo do programa é simular a capacidade humana de compreender histórias. É característico da capacidade que os seres humanos têm para compreender histórias o poderem responder a perguntas acerca destas, ainda que a informação que dêem nunca tenha ocorrido explicitamente na narrativa. Assim, por exemplo, suponha que lhe contam a seguinte história: «Um homem foi a um restaurante e pediu um hambúrguer. Quando o hambúrguer chegou estava esturricado e o homem saiu furiosamente do restaurante, sem pagar pelo hambúrguer ou deixar gorjeta.» Agora, se lhe perguntarem «O homem comeu o hambúrguer?» presumivelmente irá responder «Não, não comeu.» De modo similar, se lhe contarem a seguinte história: «Um homem entrou num restaurante e pediu um hambúrguer; quando o hambúrguer chegou o homem ficou muito agradado e ao sair do restaurante deu à empregada uma gorjeta generosa antes de pagar a conta» e lhe perguntarem: «O homem comeu o hambúrguer?» presumivelmente irá responder «Sim, ele comeu o hambúrguer.» As máquinas de Schank respondem de modo similar a perguntas acerca de restaurantes. Para fazer isto, têm uma «representação» do tipo de informação que os seres humanos têm acerca de restaurantes, o que lhes permite responder a perguntas como as mencionadas, dado este tipo de

histórias. Quando se dá a história à máquina e se lhe faz uma pergunta, a máquina imprimirá respostas do tipo que esperaríamos ouvir a seres humanos, se lhes contássemos histórias semelhantes. Os partidários da IA forte afirmam que nesta sequência de pergunta e resposta a máquina não está apenas a simular uma capacidade humana mas também 1) que podemos literalmente afirmar que a máquina *compreende* a história e fornece respostas a perguntas e 2) que a máquina e o seu programa *explicam* a capacidade humana para compreender a história e responder a perguntas acerca dela.

Não me parece que ambas as afirmações tenham de todo em todo sustentação no trabalho de Schank, como procurarei mostrar no que se segue. Não estou, obviamente, a afirmar que o próprio Schank defende estas afirmações.

Um modo de testar qualquer teoria da mente é perguntar a mim próprio como seria se a minha mente funcionasse efectivamente com base nos princípios nos quais a teoria afirma que se baseia o funcionamento de todas as mentes. Apliquemos este teste ao programa de Schank com o seguinte *Gedankenexperiment*.[*] Suponha-se que estou fechado num quarto e me dão uma sequência enorme de caracteres chineses. Suponha-se além disso (como na verdade é o caso) que não entendo seja o que for de chinês, escrito ou falado, e que não estou seguro de conseguir diferenciar a escrita chinesa de, por exemplo, escrita japonesa ou de garatujas sem sentido. Para mim, a escrita chinesa e um monte de garatujas sem sentido vão dar no mesmo. Suponha-se agora que após esta primeira sequência de caracteres chineses me dão uma segunda sequência de caracteres chineses juntamente com um conjunto de regras para correlacionar a segunda sequência com a primeira. As regras estão em português e compreendo-as tão bem como qualquer outro falante nativo do português. Elas permitem-me correlacionar um conjunto de símbolos formais com outro conjunto de símbolos formais e tudo o que «formal» significa aqui é que se pode identificar os símbolos apenas pelo seu aspecto. Suponha-se agora que também me dão uma terceira sequência de símbolos chineses juntamente com algumas instruções, novamente em português, que me permitem correlacionar elementos desta terceira sequência com as primeiras duas sequências e que estas regras me ensinam a devolver certos símbolos chineses com certos tipos de figura, em resposta a certos tipos de figura que me foram dados na terceira sequência. Sem que eu saiba, as pessoas que me dão todos estes símbolos chamam à primeira sequência um «guião», chamam à segunda sequência uma «história» e à terceira «perguntas». Além disso, chamam aos símbolos que lhes devolvo em resposta à terceira sequência «respostas às perguntas» e ao conjunto de regras em português que me deram chamam «programa». Agora, só para complicar um pouco a história, imagine-se que estas pessoas também me deram histórias em português, que compreendo, e depois me fizeram perguntas em português acerca destas histórias e que lhes dou respostas em português. Suponha-se também que, ao fim de algum tempo, me torno tão bom a seguir as instruções para manipular os símbolos chineses e os programadores se tornam tão bons a escrever os programas, que do ponto de vista externo — isto é, do ponto de vista de alguém que está fora do quarto no qual estou trancado — as minhas respostas às perguntas são absolutamente indistinguíveis das de outros falantes nativos do português, pela simples razão de que sou um falante nativo do português. Do ponto de vista externo — do ponto de vista de alguém que lê as minhas «respostas» — as respostas às perguntas em chinês e às perguntas em português são igualmente boas. Mas no caso chinês, ao contrário do português, apresento as respostas manipulando símbolos formais que não foram interpretados. No que diz respeito ao chinês, comporto-me simplesmente como um computador: executo operações computacionais sobre elementos formalmente especificados. No que diz respeito ao chinês, sou simplesmente uma instanciação de um programa de computador.

---

[*] Em alemão no original: «Experiência mental». N do T.

Ora, as afirmações que a IA forte faz são as de que o computador programado compreende as histórias e que o programa explica de algum modo o entendimento humano. Mas encontramo-nos agora em posição de examinar estas afirmações à luz da nossa experiência mental.

1 – Quanto à primeira afirmação, parece-me bastante óbvio, no exemplo, que não compreendo uma única palavra das histórias chinesas. Tenho dados de entrada e dados de saída que são indistinguíveis dos que tem um falante nativo do chinês e posso ter qualquer programa formal que se queira, mas continuo sem perceber patavina. Pelas mesmas razões, o computador de Schank nada compreende de quaisquer histórias, seja em chinês, português, ou qualquer outra língua, pois no exemplo chinês sou eu o computador e nos exemplos em que não sou o computador, este nada mais tem do que eu no exemplo em que não compreendo coisa alguma.

2 – No que respeita à segunda afirmação, a de que o programa explica o entendimento humano, podemos ver que o computador e o seu programa não fornecem condições suficientes para o entendimento uma vez que o computador e o programa funcionam sem que haja entendimento. Mas será que fornece mesmo uma condição necessária ou uma contribuição significativa para o entendimento? Uma das afirmações feitas pelos defensores da IA forte é que quando compreendo uma história em português o que faço é exactamente o mesmo — ou talvez mais do mesmo — que fazia ao manipular os símbolos chineses. O que distingue o exemplo em português, em que compreendo, do exemplo em chinês, em que não compreendo, é simplesmente uma manipulação mais formal. Não demonstrei que esta afirmação é falsa, mas pareceria seguramente uma afirmação incrível neste exemplo. A plausibilidade que a afirmação tem deriva da suposição de que podemos construir um programa que terá os mesmos dados de entrada e de saída que os falantes nativos e, além disso, pressupomos que os falantes têm algum nível de descrição em que também eles são instanciações de um programa. Com base nestas duas pressuposições presumimos que mesmo que o programa de Schank não nos dê tudo o que há para saber acerca do entendimento, poderá fazer parte disso. Bem, suponho que seja uma possibilidade empírica, mas até agora não foi apresentada a mínima razão para acreditar que seja verdadeira, uma vez que aquilo que é sugerido — embora certamente não demonstrado — pelo exemplo é que o programa de computador é simplesmente irrelevante para o meu entendimento da história. No exemplo do chinês tenho tudo aquilo que a inteligência artificial me pode dar através de um programa e nada compreendo; no exemplo do português compreendo tudo e não há até agora qualquer razão para supor que o meu entendimento tem algo a ver com programas de computador, isto é, com operações computacionais sobre elementos formalmente especificados. Enquanto o programa for definido em termos de operações computacionais sobre meros elementos formalmente definidos, o que o exemplo sugere é que por si próprios os programas não têm qualquer conexão interessante com o entendimento. Não são certamente condições suficientes e não foi dada a mínima razão para supor que são condições necessárias ou mesmo que dão um contributo significativo para o entendimento. Repare-se que a força do argumento não está apenas em que diversas máquinas podem ter os mesmos dados de entrada e de saída ao mesmo tempo que funcionam sobre princípios formais diferentes — não é essa, de todo em todo, a questão. Ao invés, sejam quais forem os princípios puramente formais que se introduza no computador, não serão suficientes para o entendimento, uma vez que um humano será capaz de seguir os princípios formais sem compreender seja o que for. Não foi apresentada qualquer razão para supor que quando compreendo o português funciono com qualquer programa formal que seja.

Então, o que terei no caso das frases portuguesas, que não tenho no caso das frases chinesas? A resposta óbvia é que sei o que as primeiras significam, ao passo que não faço a mínima ideia do que

significam as últimas. Mas em que consiste isto e por que não o poderíamos dar a uma máquina, seja o que for? Regressarei a esta questão mais tarde, mas primeiro quero continuar com o exemplo.

Já tive ocasião de apresentar este exemplo a diversas pessoas que trabalham em inteligência artificial e, curiosamente, não parecem estar de acordo sobre qual será a resposta adequada a isto. Obtenho uma variedade surpreendente de respostas e no que se segue tomarei em consideração a mais comum destas (especificadas juntamente com a sua origem geográfica).

Mas primeiro quero bloquear alguns mal-entendidos comuns acerca do «entendimento»: em muitas destas discussões encontra-se bastante jogo de cintura a propósito da palavra «entendimento». Os meus críticos sublinham que há muitos graus diferentes de entendimento; que «entendimento» não é um predicado diádico simples; que há mesmo tipos e níveis diferentes de entendimento e amiúde a lei do terceiro excluído nem sequer se aplica de um modo directo a afirmações da forma «*x* compreende *y*»; que em muitos casos é uma questão de decisão e não apenas uma questão de facto saber se *x* compreende *y*; e por aí em diante. A tudo isto quero responder: claro, claro. Mas não tem a ver com o que está em discussão. Há casos nítidos em que «entendimento» tem aplicação literal e casos nítidos em que não tem; e estes dois tipos de caso são tudo o que preciso para este argumento.[1] Compreendo histórias em português; consigo compreender, não tão bem, histórias em francês, ainda menos em alemão e em chinês nem uma palavra. O meu carro e a minha máquina de calcular, por outro lado, nada compreendem: não estão nessa linha de actividade. Atribuímos com frequência «entendimento» e outros predicados cognitivos, por metáfora e analogia, a carros, máquinas de calcular, e outros artefactos, mas estas atribuições não provam coisa alguma. Dizemos «A porta *sabe* quando abrir porque tem um sensor fotoeléctrico», «A máquina de calcular *sabe como* (*compreende como,* é *capaz de*) fazer adições e subtracções mas não divisões» e «O termóstato *percepciona* mudanças de temperatura». A razão por que fazemos estas atribuições é bastante interessante e tem a ver com o facto de estendermos a nossa própria intencionalidade aos artefactos;[2] as nossas ferramentas são extensões dos nossos objectivos de modo que achamos natural atribuir-lhes metaforicamente intencionalidade; mas julgo que tais exemplos não quebram o gelo filosófico. O sentido em que uma porta automática «compreende instruções» através do seu sensor fotoeléctrico não é de todo em todo o sentido em que compreendo o português. Se o sentido em que os computadores programados de Schank compreendem as histórias fosse supostamente o sentido metafórico em que a porta compreende e não o sentido em que compreendo o português, não valeria a pena discutir o assunto. Mas Newell e Simon (1963) afirmam que o tipo de cognição que atribuem aos computadores é exactamente o mesmo que o dos seres humanos. Gosto da franqueza desta afirmação e é o tipo de afirmação que terei em conta. Argumentarei que, no sentido literal, o computador programado compreende o mesmo que o carro e a máquina de calcular compreendem, ou seja, rigorosamente nada. O entendimento do computador não é apenas (como o meu entendimento do alemão) parcial ou incompleto: é zero.

---

[1] Além disso, «entendimento» implica quer a posse de estados mentais (intencionais) quer a verdade (validade, êxito) destes estados. Para o que interessa a esta discussão, só nos preocupamos com a posse dos estados.

[2] A intencionalidade é por definição aquela característica de certos estados mentais pela qual eles são direccionados para ou acerca de objectos e estados de coisas no mundo. Assim, crenças, desejos, intenções, são estados intencionais; formas não direccionadas de ansiedade e depressão não o são.

Passemos agora às objecções:

1) A objecção dos sistemas (Berkeley).

«Embora seja verdade que o indivíduo que está fechado no quarto não compreende a história, o facto é que ele é apenas parte de um sistema no seu todo e o sistema não compreende a história. O indivíduo tem à sua frente um registo mais amplo em que estão redigidas as regras, tem bastante papel de rascunho e lápis para fazer cálculos, tem «bancos de dados» de conjuntos de símbolos chineses. Ora, não se está a atribuir o entendimento ao mero indivíduo; ao invés, está-se a atribui-lo a todo este sistema de que o indivíduo faz parte.»

A minha resposta à teoria dos sistemas é bastante simples: permita-se ao indivíduo interiorizar todos estes elementos do sistema. Ele memoriza as regras no registo e os bancos de dados de símbolos chineses e faz todos os cálculos na sua cabeça. O indivíduo incorpora então todo o sistema. Nada há no sistema que ele não abranja. Podemos até livrar-nos do quarto e supor que ele trabalha no exterior. Ainda assim, o indivíduo nada compreende do chinês e, *a fortiori*, nem o sistema, porque nada há no sistema que não esteja no indivíduo. Se ele não compreende então não há maneira de o sistema compreender porque o sistema é apenas uma parte do indivíduo.

Na realidade, sinto-me algo envergonhado por dar mesmo esta resposta à teoria dos sistemas pois a teoria parece-me desde logo muito implausível. A ideia é a de que embora o indivíduo não compreenda o chinês, de algum modo a *combinação* desse indivíduo e dos pedaços de papel poderia compreender o chinês. Não me é fácil imaginar como poderia alguém que não estivesse preso a uma ideologia sequer considerar plausível essa ideia. Ainda assim, penso que muitas pessoas que estão comprometidas com a ideologia da IA forte se sentirão por fim inclinadas a afirmar algo de muito semelhante a isto, pelo que proponho avançarmos um pouco mais. De acordo com uma versão desta perspectiva, embora o homem no exemplo dos sistemas interiorizados não compreenda o chinês no mesmo sentido em que um falante nativo do chinês o compreende (porque, por exemplo, não sabe que a história se refere a restaurantes e hambúrgueres, etc.), ainda assim «o homem como sistema de manipulação de símbolos formais» *compreende realmente o chinês*. Não se devia confundir o subsistema do homem, que é o sistema de manipulação de símbolos formais para chinês, com o subsistema para o português.

De modo que há na verdade dois subsistemas no homem; um deles compreende o português, o outro compreende o chinês, e «acontece apenas que os dois sistemas pouco têm a ver um com o outro.» Mas, tenciono responder, não só pouco têm a ver um com o outro como nem sequer são remotamente semelhantes. O subsistema que compreende o português (pressupondo que nos permitimos temporariamente usar este jargão dos «subsistemas») sabe que as histórias são acerca de restaurantes e comer hambúrgueres, sabe que lhe fazem perguntas acerca de restaurantes e que dá respostas tão bem como pode, fazendo diversas inferências a partir do conteúdo da história, e por aí em diante. Mas o sistema chinês nada sabe disto. Ao passo que o subsistema português sabe que «hambúrgueres» refere hambúrgueres, o subsistema chinês sabe apenas que «garatuja garatuja» é seguida por «gatafunho gatafunho». Tudo o que sabe é que os diversos símbolos formais são introduzidos num lado, manipulados de acordo com regras escritas em português, e que do outro lado saem outros símbolos. O interesse do exemplo original era argumentar que tal manipulação de símbolos não podia por si ser suficiente para compreender o chinês em qualquer sentido literal, porque o homem podia escrever «gatafunho gatafunho» depois de ler «garatuja garatuja» sem perceber coisa alguma de chinês. E postular subsistemas no homem não responde ao argumento, porque os subsistemas não estão, logo à partida, melhor que o homem: continuam sem ter seja o que for de remotamente semelhante ao que o homem (ou subsistema) que é falante de português tem. Com

efeito, no exemplo descrito, o subsistema chinês é apenas uma parte do subsistema português, uma parte que se empenha numa manipulação de símbolos desprovida de sentido, feita de acordo com regras escritas em português.

Perguntemo-nos o que supostamente motiva, antes de mais, a objecção dos sistemas; isto é, que base *independente* se supõe que haja para afirmar que o agente tem de ter um subsistema no seu interior, que compreende literalmente as histórias em chinês? Tanto quanto vejo, a única base é a de que, no exemplo, tenho os mesmos dados de entrada e de saída que os falantes nativos de chinês e um programa que vai de um a outro. Mas o propósito dos exemplos tem sido o de tentar mostrar que isso não podia ser suficiente para o entendimento, no sentido em que compreendo histórias em português, porque uma pessoa, e portanto um conjunto de sistemas que constitui uma pessoa, podia ter a combinação certa de dados de entrada, dados de saída e um programa e ainda assim não compreender coisa alguma no sentido relevante, literal, em que compreendo o português. A única motivação para afirmar que *tem* de haver um subsistema em mim que compreende o chinês é que tenho um programa e posso passar o teste de Turing; posso enganar os falantes nativos do chinês. Mas a adequação do teste de Turing é precisamente um dos pontos em debate. O exemplo mostra que podia haver dois «sistemas», de tal modo que ambos passam o teste de Turing mas apenas um compreende; e contra isto não serve de argumento afirmar que uma vez que ambos passam o teste de Turing ambos têm de compreender, uma vez que esta afirmação não é capaz de responder ao argumento de que o sistema em mim que compreende o português tem muito mais que o sistema que meramente processa o chinês. Resumindo, a objecção dos sistemas cai simplesmente em petição de princípio ao insistir, sem argumentação, que os sistemas têm de compreender o chinês.

Além disso, a objecção dos sistemas parece levar a consequências que são independentemente absurdas. Se concluirmos que tem de haver cognição em mim com base em que tenho um certo tipo de dados de entrada e de saída e um programa entre ambos, parece então que se acabará por considerar cognitivos todo o tipo de subsistemas que não o são. Por exemplo, há um nível de descrição no qual o meu estômago processa informação e instancia quaisquer programas de computador, mas presumo que não queiramos afirmar que o meu estômago compreende seja o que for (cf. Pylyshyn, 1980). Mas se aceitamos a objecção dos sistemas, é difícil ver como evitamos afirmar que o estômago, o coração, o fígado, e por aí em diante, são todos subsistemas com entendimento, uma vez que não há meio privilegiado de distinguir a motivação para afirmar que o subsistema chinês compreende da de afirmar que o estômago compreende. Por sinal, não se responde a isto afirmando que o sistema chinês tem informação como dados de entrada e de saída e que o estômago tem alimentos como dados de entrada e alimentos digeridos como dados de saída, uma vez que do ponto de vista do agente, do meu ponto de vista, não há informação quer na comida quer nos caracteres chineses — o chinês equivale aqui a garatujas sem sentido. A informação, no exemplo do chinês, está apenas no olhar dos programadores e dos intérpretes, e nada os impede de tratar os dados de entrada e de saída dos meus órgãos digestivos como informação, se assim o desejarem.

Este último ponto influi em alguns problemas independentes da IA forte e vale a pena divagar por um momento para o explicar. Se a IA forte faz parte da psicologia, então tem de ser capaz de distinguir os sistemas que são genuinamente mentais dos que não o são. Tem de ser capaz de distinguir os princípios sobre os quais funciona a mente daqueles sobre os quais funcionam os sistemas que não são mentais; de contrário não nos oferecerá explicações acerca do que é especificamente mental acerca do mental. E a distinção mental-amental não pode apenas estar no olhar do observador mas tem de ser intrínseca aos sistemas; de contrário caberia a cada observador tratar as pessoas como amentais e, por exemplo, os ciclones como mentais se lhe aprouvesse. Mas na

bibliografia da IA a distinção é com muita frequência esbatida de maneiras que a longo prazo se mostrariam desastrosas para a afirmação de que a IA é um inquérito cognitivo. McCarthy, por exemplo, afirma: «Pode dizer-se que as máquinas simples como os termóstatos têm crenças e ter crenças parece ser uma característica, na sua maioria, das máquinas capazes de resolver problemas.» (McCarthy, 1979). Quem quer que pense que a IA forte tem alguma hipótese como teoria da mente devia ponderar as implicações deste comentário. Convidam-nos a aceitá-lo como uma descoberta da IA forte, que o pedaço de metal na parede, que usamos para regular a temperatura, tem crenças exactamente no mesmo sentido que nós, os nossos cônjuges, as nossas crianças, temos crenças e que além disso, «na sua maioria», as outras máquinas que estão no quarto — o telefone, o gravador de cassetes, a máquina de calcular, o electric fight switch — também têm crenças neste sentido literal. O objectivo deste artigo não é argumentar contra o ponto de vista de McCarthy, pelo que irei simplesmente afirmar o que se segue sem argumento. O estudo da mente começa com factos como: que os humanos têm crenças, ao passo que os termóstatos, telefones e máquinas de calcular não têm. Se temos uma teoria que nega este ponto, produzimos um contra-exemplo à teoria e a teoria é falsa. Ficamos com a impressão de que as pessoas envolvidas na IA que escrevem este tipo de coisa pensam poder safar-se com isso porque não o levam realmente a sério e também não pensam que mais alguém o fará. Proponho, ao menos por um momento, que o levemos a sério. Pensemos arduamente por um minuto sobre o que seria necessário para estabelecer que aquele pedaço de metal na parede ali tem crenças a sério, crenças com direcção de adequação, conteúdo proposicional, condições de satisfação; crenças com a possibilidade de ser fortes ou fracas; crenças nervosas, ansiosas ou seguras; dogmáticas, racionais ou supersticiosas; fé cega ou cogitações hesitantes, qualquer tipo de crenças. O termóstato não é um candidato. Tão-pouco o é o estômago, o fígado, a máquina de calcular, o telefone. Contudo, uma vez que tomamos a ideia a sério, repare-se que a sua verdade seria fatal para a pretensão da IA forte em ser uma ciência da mente. Pois agora a mente está em todo o lado. O que queríamos saber é o que distingue a mente dos termóstatos e fígados. E se McCarthy tivesse razão, a IA forte não teria a mais leve esperança de nos dizer o que é.

2) A Objecção do Robô (Yale).

«Suponha-se que escrevíamos um tipo de programa diferente do de Schank. Suponha-se que colocamos um computador dentro de um robô e que este computador não se limitaria a aceitar símbolos formais como dados de entrada e a emitir símbolos formais como dados de saída, mas que, ao invés, manobraria realmente o robô de tal modo que este faria algo muito semelhante a percepcionar, caminhar, mover-se de um lado para o outro, pregar pregos, comer, beber — o que se queira. O robô teria, por exemplo, uma câmara de vídeo integrada que lhe permitiria ver, teria braços e pernas que lhe permitiriam «agir» e tudo isto seria controlado pelo seu «cérebro» computorizado. Tal robô, ao contrário do computador de Schank, teria entendimento genuíno e outros estados mentais.»

A primeira coisa a notar acerca da objecção do robô é que concede tacitamente que a cognição não é apenas uma questão de manipular símbolos formais, dado que esta resposta acrescenta um conjunto de relações causais com o mundo exterior (cf. Fodor, 1980). Mas a resposta à objecção do robô é que a adição de tais capacidades «perceptivas» e «motoras» nada acrescenta a propósito do entendimento em particular ou da intencionalidade em geral, ao programa original de Schank. Para ver isto, repare-se que a mesma experiência mental se aplica ao caso do robô. Suponha-se que em vez do computador dentro do robô, me colocam a mim dentro do quarto e, como no exemplo original do chinês, me dão mais símbolos chineses com mais instruções em português para combinar símbolos chineses com símbolos chineses e reenviar símbolos chineses para o exterior. Suponha-se que, sem eu saber, alguns

dos símbolos chineses que me fazem chegar vêm de uma câmara de vídeo integrada no robô e que outros símbolos chineses que faço sair servem para fazer que os motores instalados no interior do robô lhe movam as pernas ou os braços. É importante sublinhar que tudo o que faço é manipular símbolos formais: nada sei acerca dos outros factos. Recebo «informação» vinda do equipamento «perceptivo» do robô e emito «instruções» ao seu equipamento motor sem ter conhecimento de qualquer destes factos. Sou o homúnculo do robô, mas ao contrário do homúnculo tradicional, não sei o que se passa. Nada compreendo excepto as regras para a manipulação de símbolos. Neste caso quero dizer que o robô não tem quaisquer estados intencionais de todo em todo; apenas se move de um lado para o outro em resultado dos seus circuitos eléctricos e do seu programa. E além disso, ao instanciar o programa não tenho quaisquer estados intencionais do tipo relevante. Tudo o que faço é seguir instruções formais acerca da manipulação de símbolos formais.

3) A objecção do simulador de cérebros (Berkeley e MIT).

«Suponha-se que concebemos um programa que não representa a informação que temos acerca do mundo, como a informação nas sequências de caracteres de Schank, mas que simula a sequência efectiva do disparar de neurónios nas sinapses do cérebro de um falante nativo de chinês quando este compreende histórias em chinês e responde a perguntas acerca delas. A máquina recebe histórias e perguntas acerca destas como dados de entrada, simula a estrutura formal de cérebros chineses efectivos ao processar estas histórias e emite respostas em chinês como dados de saída. Podemos até imaginar que a máquina não funciona com um único programa serial mas com todo um conjunto de programas funcionando em paralelo, do modo como presumivelmente funcionam os cérebros humanos efectivos quando processam a linguagem natural. Agora, seguramente que nesse caso teríamos de dizer que a máquina compreende as histórias. E, se nos recusarmos a dizê-lo, não teríamos também de negar que os falantes nativos do chinês compreendem as histórias? Ao nível das sinapses, o que seria ou poderia ser diferente no programa de computador e no programa do cérebro chinês?»

Antes de rebater esta objecção quero fazer um desvio para notar que se trata de uma objecção bizarra para qualquer partidário da inteligência artificial (ou funcionalismo, etc.): pensei que a ideia da IA forte fosse a de não precisar de saber como o cérebro funciona para saber como a mente funciona. A hipótese básica, ou assim supus, era a de que há um nível das operações mentais que consiste em processos computacionais sobre elementos formais que constituem a essência do mental e podem ser concretizados em todo o tipo de processos cerebrais diferentes, do mesmo modo que qualquer programa de computador pode ser concretizado em diferentes equipamentos informáticos: com os pressupostos da IA forte, a mente é para o cérebro o que o programa é para o *hardware* e assim podemos compreender a mente sem fazer neurofisiologia. Se tivéssemos de saber como o cérebro funcionou para fazer a IA não teríamos de nos preocupar com a IA. Contudo, mesmo aproximando isto do funcionamento do cérebro não é ainda suficiente para produzir o entendimento. Para ver isto, imagine-se que ao invés de um homem monolingue a reordenar símbolos dentro de um quarto temos o homem a manobrar um conjunto elaborado de tubagens de água, com válvulas a conectá-las. Quando o homem recebe os símbolos chineses vai ver ao programa, escrito em português, que válvulas tem de ligar e desligar. Cada ligação na tubagem de água corresponde a uma sinapse no cérebro chinês e todo o sistema está instalado de modo a que depois de iniciar todos os disparos correctos, isto é, depois de ligar todas as torneiras correctas, as respostas chinesas surgem na porta de saída da série de tubos.

Onde está o entendimento neste sistema? Recebe chinês como dados de entrada, simula a estrutura formal das sinapses de um cérebro chinês e devolve chinês como dados de saída. Mas decerto nem o homem nem a tubagem de água compreendem o chinês e se nos sentimos tentados a adoptar o que julgo ser a perspectiva absurda de que, de algum modo, a *combinação* de homem *e* tubagem compreende, lembremo-nos de que em princípio o homem pode interiorizar a estrutura formal da tubagem e fazer todos os «disparos neurais» na sua imaginação. O problema com o simulador de cérebros é o simular as coisas erradas a respeito do cérebro. Enquanto apenas simular a estrutura formal da sequência de disparos neurais nas sinapses, não terá simulado o que importa a respeito do cérebro, nomeadamente, as propriedades causais, a sua capacidade de produzir estados intencionais. E que as propriedades formais não são suficientes para as propriedades causais vê-se pelo exemplo da tubagem: podemos ter todas as propriedades formais destacadas das propriedades causais, neurobiológicas, relevantes.

John Smith 08/4/15 23:15
Comment: Carved off

4) A objecção da combinação (Berkeley e Stanford).

«Ao passo que cada uma das três objecções anteriores pode não ser inteiramente convincente em si como refutação do contra-exemplo do quarto chinês, as três tomadas em conjunto são colectivamente mais convincentes e mesmo decisivas. Imagine-se um robô com um computador em forma de cérebro alojado na sua cavidade craniana, imagine-se o computador programado com todas as sinapses de um cérebro humano, imagine-se que todo o comportamento do robô é indistinguível do comportamento humano e agora pense-se em tudo isso como um sistema unificado e não apenas como um computador com dados de entrada e de saída. Certamente que em tal caso teríamos de atribuir intencionalidade ao sistema.»

Concordo inteiramente que em tal caso acharíamos racional e de facto irresistível aceitar a hipótese de que o robô tem intencionalidade, desde que nada mais soubéssemos acerca do mesmo. De facto, para além da aparência e do comportamento, os outros elementos da combinação são na verdade irrelevantes. Se pudéssemos construir um robô cujo comportamento fosse indistinguível de um vasto leque de comportamentos humanos, atribuir-lhe-íamos intencionalidade, até termos alguma razão para não o fazer. Não precisaríamos de saber antecipadamente que o seu cérebro computorizado era um análogo formal do cérebro humano.

Mas não creio realmente que isto ajude seja no que for as afirmações da IA forte e eis porquê: de acordo com a IA forte, instanciar um programa formal com os dados de entrada e de saída correctos é uma condição suficiente e deveras constitutiva da intencionalidade. Nas palavras de Newell (1979), a essência do mental é o funcionamento de um sistema de símbolos formais. Mas as atribuições de intencionalidade que fazemos ao robô neste exemplo nada têm a ver com programas formais. Baseiam-se simplesmente na pressuposição de que se o robô se parece e se comporta suficientemente como nós, então suporíamos, até se provar o contrário, que tem de ter estados mentais como os nossos, que causam o seu comportamento e neste se manifestam, e que tem de ter um mecanismo interno capaz de produzir tais estados mentais. Se soubéssemos como explicar independentemente o seu comportamento sem tais pressuposições, não lhe atribuiríamos a intencionalidade, especialmente se soubéssemos que tinha um programa formal. E isto é precisamente o que quero dizer com a minha resposta anterior à objecção 2).

Suponhamos que sabíamos que o comportamento do robô se explicava inteiramente pelo facto de que um homem colocado no seu interior recebia símbolos formais ininterpretados a partir dos receptores sensoriais do robô e emitia símbolos formais aos seus mecanismos motores e que o homem

fazia esta manipulação de símbolos de acordo com um punhado de regras. Além disso, suponhamos que o homem desconhece todos estes factos acerca do robô, tudo o que sabe é que operações executar sobre que símbolos desprovidos de sentido. Nesse caso veríamos o robô como um engenhoso boneco mecânico. A hipótese de que o boneco tem uma mente seria agora injustificada e desnecessária, pois não há mais razão alguma para atribuir intencionalidade ao robô, ou ao sistema de que faz parte, (excepto, claro, pela intencionalidade do homem ao manipular os símbolos). As manipulações de símbolos formais prosseguem, os dados de entrada e de saída são correctamente articulados, mas o único verdadeiro *locus* da intencionalidade é o homem e ele desconhece quaisquer dos estados intencionais relevantes; não *vê*, por exemplo, o que chega aos olhos do robô, não *intenta* mover o braço do robô e não *compreende* quaisquer dos comentários feitos para ou pelo robô. Tão-pouco, pelas razões apresentadas antes, é o sistema de que fazem parte o homem e o robô que faz estas coisas.

Para ver este ponto, contraste-se o exemplo com outros em que achamos inteiramente natural atribuir intencionalidade a membros de determinadas espécies de primatas, como o chimpanzé e o babuíno,[*] e a animais domésticos, como os cães. As razões que consideramos naturais são, *grosso modo*, duas: não podemos compreender o comportamento do animal sem a atribuição de intencionalidade e podemos ver que a matéria de que são feitos os bichos é similar à de que somos feitos — isto é um olho, aquilo um nariz, isto é a sua pele, e por aí em diante. Dada a coerência do comportamento do animal e o pressuposto de que lhe subjaz a mesma matéria causal, pressupomos que o animal tem de ter estados mentais subjacentes ao seu comportamento e que os estados mentais têm de ser produzidos por mecanismos feitos de uma matéria semelhante à de que somos feitos. Faríamos certamente pressuposições semelhantes acerca do robô a menos que tivéssemos alguma razão para não o fazer, mas assim que soubéssemos que o comportamento era o resultado de um programa formal e que as propriedades causais efectivas da substância física eram irrelevantes abandonaríamos a pressuposição de intencionalidade.

---

[*] No original: «... other primate species such as apes and monkeys...». A distinção aqui é entre o *símio* (primatas sem cauda: chimpanzé, gorila, orangotango) e o *macaco* (primatas com cauda: babuíno, saguim, lémure). Uma vez que é comum traduzir «ape» e «monkey» indiferentemente por «macaco», a opção de usar respectivamente «chimpanzé» e «babuíno» justifica-se por tornar o texto mais claro. N do T.

Há duas outras respostas ao meu exemplo que surgem frequentemente (de modo que vale a pena discuti-las) mas que na verdade passam ao lado da questão.

5) A objecção das outras mentes (Yale).

«Como sabe que as outras pessoas compreendem o chinês ou outra coisa qualquer? Apenas pelo seu comportamento. Ora o computador pode passar nos testes comportamentais tão bem como as pessoas (em princípio), de modo que se vamos atribuir cognição a outras pessoas temos em princípio de a atribuir também aos computadores.»

Esta objecção merece na verdade apenas uma resposta curta. O problema nesta discussão não é acerca de como sei que as outras pessoas têm estados cognitivos mas antes o que lhes estou a atribuir quando lhes atribuo estados cognitivos. O impulso do argumento é que não se podia tratar apenas de processos computacionais e dos seus dados de saída porque pode haver os processos computacionais e os dados de saída sem que haja o estado cognitivo. Não é resposta a este argumento fingir anestesia.

Nas «ciências cognitivas» pressupõe-se a realidade e cognoscibilidade do mental do mesmo modo que nas ciências físicas se tem de pressupor a realidade e cognoscibilidade dos objectos físicos.

6) A objecção das muitas divisórias (Yale).

John Smith 08/4/16 10:50
Comment: Many mansions

«Todo o seu argumento pressupõe que a IA é apenas acerca de computadores analógicos e digitais. Mas acontece apenas que isso corresponde ao presente estado da tecnologia. Sejam o que forem estes processos causais que diz serem essenciais para a intencionalidade (pressupondo que está correcto), seremos eventualmente capazes de construir aparelhos que tenham estes processos causais e que serão inteligência artificial. Pelo que os seus argumentos não se dirigem de modo algum à capacidade da inteligência artificial em produzir e explicar a cognição.»

Nada tenho realmente a objectar a esta objecção excepto dizer que, com efeito, trivializa o projecto da IA forte redefinindo-o como seja o que for que produza artificialmente a cognição e a explique. O interesse da afirmação original feita em nome da inteligência artificial é o de ser uma tese precisa e bem definida: os processos mentais são processos computacionais sobre elementos definidos formalmente. Desafiar esta tese tem sido a minha preocupação. Se a afirmação é redefinida de tal modo que deixa de ser essa tese, as minhas objecções já não se aplicam, porque deixou de haver uma hipótese testável à qual se apliquem.

Regressemos agora à questão que prometi tentar responder: admitindo que no meu exemplo original compreendo o português e não compreendo o chinês e admitindo-se portanto que a máquina não compreende nem o português nem o chinês, ainda assim tem de haver em mim algo que faz que seja o caso que eu compreendo o português e uma correspondente ausência em mim de algo que faz que seja o caso que não consigo compreender o chinês. Ora, por que não poderíamos dar essas coisas, sejam elas o que forem, a uma máquina?

Não vejo qualquer razão por que não poderíamos em princípio dar a uma máquina a capacidade de compreender o português ou o chinês, visto que, num sentido importante, os nossos corpos, com os nossos cérebros, são precisamente tais máquinas. Mas vejo de facto argumentos muito importantes para afirmar que não poderíamos dar tal coisa a uma máquina cujo funcionamento se defina apenas em termos de processos computacionais sobre elementos formalmente definidos; isto é, em que o funcionamento da máquina é definido como uma instanciação de um programa de computador. Não é por ser a instanciação de um programa de computador que sou capaz de compreender o português e de ter outras formas de intencionalidade (Sou, suponho, a instanciação de quaisquer programas de computador) mas, tanto quanto sabemos, por pertencer a um certo tipo de organismo com uma certa estrutura biológica (isto é, química e física) e esta estrutura, sob certas condições, é causalmente capaz de produzir a percepção, acção, entendimento, aprendizagem e outros fenómenos intencionais. E em parte a ideia do presente argumento é a de que apenas algo que tivesse esses poderes causais poderia ter intencionalidade. Talvez outros processos físicos e químicos pudessem produzir exactamente estes efeitos; talvez, por exemplo, os marcianos também tenham intencionalidade ainda que os seus cérebros sejam feitos de uma matéria diferente. Essa é uma questão empírica, semelhante à questão de saber se a fotossíntese pode ser feita por algo com uma constituição química diferente da que tem a clorofila.

Mas o ponto principal do presente argumento é que nenhum modelo puramente formal será alguma vez suficiente em si para a intencionalidade porque as propriedades formais não são por si próprias constitutivas da intencionalidade e não têm por si próprias quaisquer poderes causais excepto o poder,

quando instanciado, de produzir a próxima etapa do formalismo quando a máquina está em funcionamento. E quaisquer outras propriedades causais que tenham concretizações particulares do modelo formal são irrelevantes para o modelo formal porque podemos sempre pôr o mesmo modelo formal numa concretização diferente, em que essas propriedades causais estão obviamente ausentes. Mesmo se, por algum milagre, os falantes do chinês concretizam exactamente o programa de Schank, podemos pôr o mesmo programa em falantes do português, tubagens, ou computadores, sendo que nenhum destes compreende o chinês, apesar do programa.

O que importa acerca das operações cerebrais não é a sombra formal projectada pela sequência de sinapses mas antes as propriedades efectivas das sequências. Todos os argumentos em favor da versão forte da inteligência artificial que tenho visto insistem em traçar um contorno em redor das sombras projectadas pela cognição afirmando depois que as sombras são o produto genuíno.

Em jeito de conclusão, quero tentar afirmar algumas das questões filosóficas gerais implícitas no argumento. Em abono da clareza, procurarei fazê-lo através da pergunta e resposta e começo por aquela já velha pergunta:

«Poderia uma máquina pensar?» A resposta é, obviamente, sim. Somos precisamente tais máquinas.

«Sim, mas poderia um artefacto, uma máquina feita pelo homem, pensar?»

Pressupondo que é possível produzir artificialmente uma máquina com sistema nervoso, neurónios com axónios e dendrites e tudo o mais, suficientemente parecido com o nosso, mais uma vez a resposta parece ser, obviamente, sim. Se podemos duplicar exactamente as causas, podemos duplicar os efeitos. E de facto poderia ser possível produzir a consciência, a intencionalidade, e tudo o mais, usando quaisquer outros tipos de princípios químicos que não os que constituem os seres humanos. É, como disse, uma questão empírica.

«Ok, mas poderia um computador digital pensar?»

Se por «computador digital» entendemos qualquer coisa que tenha um nível de descrição em que possa ser correctamente descrita como a instanciação de um programa de computador, então a resposta é uma vez mais, obviamente, sim, uma vez que somos a instanciação de quaisquer programas de computador e podemos pensar.

«Mas poderia algo pensar, compreender, e por aí em diante, *apenas* em virtude de ser um computador com o tipo correcto de programa? Poderia a instanciação de um programa, o programa correcto, claro, ser em si uma condição suficiente para o entendimento?»

Penso que esta é a questão certa a colocar, embora normalmente se confunda com uma ou outra das questões anteriores, e a resposta é não.

«Por que não?»

Porque as manipulações de símbolos formais por si próprias não têm qualquer intencionalidade; são inteiramente desprovidas de sentido; não são sequer manipulações de *símbolos*, uma vez que os símbolos não simbolizam coisa alguma. No jargão linguístico, têm apenas uma sintaxe mas nenhuma

semântica. Tal intencionalidade que os computadores aparentam ter está apenas nas mentes dos que os programam e nas dos que os usam, que são quem fornece os dados de entrada e interpreta os dados de saída.

O objectivo do exemplo do quarto chinês era tentar mostrar isto mostrando que assim que colocamos no sistema algo que tenha realmente intencionalidade (um homem) e o programamos com o programa formal, pode-se ver que o programa formal não traz qualquer intencionalidade adicional. Nada acrescenta, por exemplo, à capacidade que o homem tem de compreender o chinês.

Precisamente aquela característica da IA que parecia tão apelativa — a distinção entre o programa e a concretização — mostra-se fatal para a afirmação de que simulação podia ser duplicação. A distinção entre o programa e a sua concretização no *hardware* parece ser paralela à distinção entre o nível das operações mentais e o nível das operações cerebrais. E se pudéssemos descrever o nível das operações mentais como um programa formal, então parece que podíamos descrever o que seria essencial acerca da mente sem fazer quer psicologia introspectiva quer neurofisiologia do cérebro. Mas a equação «a mente é para o cérebro o que o programa é para o *hardware*» soçobra em vários pontos, entre eles os seguintes três:

Em primeiro lugar, a distinção entre programa e realização tem a consequência de que o mesmo programa podia ter todo o tipo de realizações malucas que não tivessem qualquer forma de intencionalidade. Weizenbaum (1976, Cap. 2), por exemplo, mostra em detalhe como construir um computador usando um rolo de papel higiénico e uma pilha de pequenos calhaus. Similarmente, o programa para compreender a narrativa chinesa pode ser programado numa sequência de tubagens, num conjunto de ventoinhas, ou num falante monolingue de português, nenhum dos quais obtém uma compreensão do chinês. Pedras, papel higiénico, vento, tubagens, são desde logo o tipo inadequado de coisa para se obter a intencionalidade — só algo que tenha os mesmos poderes causais que os cérebros pode ter intencionalidade — e embora o falante de português tenha o tipo adequado de matéria para a intencionalidade pode-se ver facilmente que este não obtém qualquer intencionalidade adicional por memorizar o programa, uma vez que memorizar não lhe ensinará o chinês.

Em segundo lugar, o programa é puramente formal, mas os estados intencionais não são formais desse modo. Definem-se em termos do seu conteúdo, não da sua forma. A crença de que está a chover, por exemplo, não se define como um certo aspecto formal mas como um certo conteúdo mental com condições de satisfação, direcção de adequação (ver Searle, 1979), e coisas semelhantes. Com efeito, a forma como tal não tem sequer um aspecto formal neste sentido sintáctico, uma vez que se pode dar a uma única crença um número indefinido de expressões sintácticas diferentes em sistemas linguísticos diferentes.

John Smith 08/4/16 00:16
Comment: Formal shape

Em terceiro lugar, como mencionei antes, os estados e acontecimentos mentais são literalmente produtos do funcionamento do cérebro, mas o programa não é, do mesmo modo, um produto do computador.

«Bem, se os programas não são, de modo algum, constitutivos dos processos mentais, por que razão tem muita gente acreditado no contrário? Isso pelo menos pede alguma explicação.»

Não sei mesmo qual será a resposta a isto. A ideia de que as simulações de computador pudessem ser o produto genuíno deve ter parecido à partida suspeita porque o computador não está confinado, por quaisquer meios, à simulação de operações mentais. Ninguém supõe que as simulações

computorizadas de um incêndio de grau 5 irão destruir a vizinhança ou que uma simulação de computador de uma chuva torrencial nos fará ficar encharcados. Por que raios iria alguém supor que uma simulação computorizada do entendimento tem com efeito entendimento acerca de qualquer coisa? Diz-se por vezes que seria assustadoramente difícil fazer que os computadores sintam dor ou se apaixonem, mas o amor e a dor não são nem mais difíceis nem mais fáceis que a cognição ou seja o que for. Para a simulação, tudo o que é preciso são dados de entrada e dados de saída e um programa no meio, que transforma os primeiros nos segundos. Isso é tudo o que o computador tem para tudo o que faz. Confundir a simulação com a duplicação é fazer o mesmo erro, quer se trate de dor, amor, cognição, incêndios, ou tempestades.

Ainda assim, há diversas razões por que tem de ter parecido — e a muita gente talvez ainda pareça — que de algum modo a IA reproduz e portanto explica os fenómenos mentais e acredito que não conseguiremos remover estas ilusões antes de ter exposto integralmente as razões que lhes dão origem.

Em primeiro lugar e sendo talvez o mais importante, está uma confusão acerca da noção de «processamento de informação»: muita gente nas ciências cognitivas acredita que o cérebro humano, com a sua mente, faz algo a que se chama «processamento de informação» e analogamente o computador com o seu programa faz processamento de informação; mas os incêndios e as tempestades, por outro lado, não fazem qualquer processamento de informação. Assim, embora o computador possa simular as características formais de qualquer processo que seja, encontra-se numa relação especial com a mente e o cérebro porque quando o computador é adequadamente programado, idealmente com o mesmo programa que o cérebro, o processamento da informação é idêntico nos dois casos e este processamento de informação é realmente a essência do mental. Mas o problema com este argumento é repousar numa ambiguidade na noção de «informação». No sentido em que as pessoas «processam informação» quando reflectem, por exemplo, em problemas de aritmética, ou quando lêem histórias e respondem a perguntas acerca destas, o computador não faz «processamento de informação». Ao invés, o que faz é manipular símbolos formais. O facto de o programador e o intérprete dos dados de saída do computador usarem os símbolos para referir objectos no mundo supera completamente o alcance do computador. O computador, repetindo o que foi dito, tem sintaxe mas não tem semântica. Assim, se alguém teclar no computador «2 + 2 =?» este responderá «4». Mas não tem ideia de que «4» significa 4 ou que significa seja o que for. E a questão não é que lhe falte alguma informação de segunda ordem acerca da interpretação dos seus símbolos de primeira ordem, mas antes que os seus símbolos de primeira ordem não têm quaisquer interpretações no que diz respeito ao computador. Tudo o que o computador tem são mais símbolos. A introdução da noção de «processamento de informação» produz portanto um dilema: ou interpretamos a noção de «processamento de informação» de tal modo que implique a intencionalidade como parte do processo ou não o fazemos. No primeiro caso, então o computador programado não faz processamento de informação, apenas manipula símbolos formais. No segundo caso, então, apesar de o computador fazer processamento de informação, só o faz no sentido em que as máquinas de calcular, as máquinas de escrever, os termóstatos, as tempestades, os ciclones, fazem processamento de informação; nomeadamente, têm um nível de descrição em que se os pode descrever como recebendo informação de um lado, transformando-a, e produzindo informação como dados de saída. Mas neste caso cabe aos observadores externos interpretar os dados de entrada e de saída como informação no sentido corrente. E não se estabelece qualquer semelhança entre o computador e o cérebro em termos de qualquer semelhança de processamento de informação.

Em segundo lugar, há um behaviourismo ou operacionalismo residuais em muito da IA. Uma vez que os computadores adequadamente programados podem ter padrões de *input-output*[*] similares aos dos seres humanos, sentimo-nos tentados a postular estados mentais no computador, similares aos estados mentais humanos. Mas assim que se veja que é conceptual e empiricamente possível que um sistema tenha capacidades humanas em algum domínio sem ter intencionalidade de todo em todo, devíamos ser capazes de superar este impulso. A minha calculadora de secretária tem capacidades de cálculo mas nenhuma intencionalidade e neste artigo tentei mostrar que um sistema podia ter capacidades relativas a dados de entrada e dados de saída que duplicassem as de um falante nativo do chinês e ainda assim não compreender o chinês, a despeito do modo como foi programado. O teste de Turing é típico da tradição de ser-se descaradamente behaviourista e operacionalista e acredito que se as pessoas que trabalham em IA repudiassem o behaviourismo e o operacionalismo, muita da confusão entre a simulação e a duplicação seria eliminada.

[*] Ao longo do texto, tem-se traduzido «input» e «output» por «dados de entrada» e «dados de saída». Optou-se neste caso por manter os termos originais para assim obter uma expressão mais simples e clara. N do T.

Em terceiro lugar, este operacionalismo residual junta-se a uma forma residual de dualismo; com efeito, a IA forte só faz sentido dada a pressuposição dualista de que, no que diz respeito à mente, o cérebro não importa. Na IA forte (e também no funcionalismo) o que importa são os programas e os programas são independentes da sua concretização nas máquinas; com efeito, no que diz respeito à IA, o mesmo programa podia ser concretizado por uma máquina electrónica, uma substância mental cartesiana, ou um hegeliano espírito do mundo. A única descoberta surpreendente que fiz ao discutir estes assuntos é que muitas pessoas que trabalham em IA se sentem muito chocadas pela minha ideia de que os efectivos fenómenos mentais humanos podem ser dependentes de efectivas propriedades físico-químicas dos cérebros humanos efectivos. Mas se pensarmos nisto durante um minuto podemos ver que eu não devia ter ficado surpreendido; pois a menos que se aceite alguma forma de dualismo, o projecto da IA forte não tem qualquer hipótese. O projecto consiste em reproduzir e explicar o mental através do *design* de programas, mas a menos que a mente seja não só conceptualmente mas também empiricamente independente do cérebro, o projecto não poderia ser levado a cabo, pois o programa é completamente independente de qualquer concretização. A menos que se acredite que a mente é separável do cérebro quer conceptualmente quer empiricamente — dualismo num sentido forte — não se pode esperar reproduzir o mental escrevendo e executando programas, uma vez que os programas têm de ser independentes dos cérebros ou de quaisquer outras formas particulares de instanciação. Se as operações mentais consistem em operações computacionais sobre símbolos formais, então segue-se que não têm qualquer conexão interessante com o cérebro; a única conexão seria que o cérebro, por acaso, é um dos indefinidamente muitos tipos de máquinas capazes de instanciar o programa. Esta forma de dualismo não é a tradicional variedade cartesiana, que afirma que há dois tipos de *substância*, mas é cartesiana no sentido de que insiste em que o que é especificamente mental acerca da mente não tem conexão intrínseca com as propriedades efectivas do cérebro. Este dualismo subjacente é disfarçado pelo facto de a bibliografia da IA conter amiúde tiradas fulminantes contra o «dualismo»; aquilo de que os autores parecem não estar cientes é que a sua posição pressupõe uma versão forte do dualismo. «Poderia uma máquina pensar?» A minha própria perspectiva é a de que *apenas* uma máquina pode pensar e com efeito apenas tipos muito especiais de máquinas, nomeadamente cérebros e máquinas que tenham os mesmos poderes causais que os cérebros. E essa é a principal razão por que a IA forte tem pouco para nos dizer acerca do pensamento, uma vez que nada tem para nos dizer acerca das máquinas. Pela sua própria definição, é acerca de programas e os programas não são máquinas. O mais que a intencionalidade seja, é um fenómeno biológico e é

provável que seja tão causalmente dependente da bioquímica específica das suas origens como a lactação, a fotossíntese, ou quaisquer outros fenómenos biológicos. Ninguém suporia que podemos produzir leite ou açúcar executando uma simulação computorizada das sequências formais na lactação e fotossíntese, mas no que diz respeito à mente muitas pessoas estão dispostas a acreditar em tal milagre por causa de um profundo e duradouro dualismo: supõem que a mente é uma questão de processos formais e é independente das causas materiais bastante específicas de um modo que o leite e o açúcar não são. Em defesa deste dualismo exprime-se amiúde a esperança de que o cérebro seja um computador digital. Uma vez que tudo são computadores digitais, os cérebros também o são. A questão é que a capacidade causal que o cérebro tem para produzir a intencionalidade não pode consistir no facto de instanciar um programa de computador, uma vez que para qualquer programa que se queira é possível que algo instancie esse programa e ainda assim não ter quaisquer estados mentais. Seja o que for que o cérebro faz para produzir a intencionalidade, isso não pode consistir na instanciação de um programa visto que nenhum programa, por si próprio, é suficiente para a intencionalidade.[3]

[3] Estou em dívida para com um número bastante elevado de pessoas pela discussão destes assuntos e pelas suas tentativas pacientes de superar a minha ignorância em inteligência artificial. Gostaria em especial de agradecer a Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, e Terry Winograd.