Cérebros numa cuba *

Brains in a Vat

Hilary Putnam

Tradução de L. H. Marques Segundo
Universidade Federal de Santa Catarina

Uma formiga passa por cima de um pequeno monte de areia. Na medida em que vai passando, traça nela uma linha. Por puro acaso, a linha faz curvas e se cruza de tal modo que claramente acaba parecendo com uma caricatura de Winston Churchill. Terá a formiga traçado uma imagem de Winston Churchill, uma imagem que *retrate* Churchill? Muitos, à primeira vista, diriam que não. Afinal, a formiga nunca vira Churchill, ou sequer uma imagem de Churchill, e não tivera a intenção de retratar Churchill. Ela simplesmente traçou uma linha (não intencionalmente), uma linha que *nós* conseguimos "ver como" uma imagem de Churchill.

Podemos expressar isso dizendo que a linha "em si" não é uma representação[2] de

* "Brains in a Vat", in *Reason, Truth, and History*, pp. 1-21.

2 - Neste [artigo] os termos "representação" e "referência" vão se referir sempre à relação entre uma palavra (ou outro tipo de signo, símbolo, ou representação) e algo que efetivamente existe (*i.e.* não apenas a "objeto do pensamento"). Há um sentido de "referir" no qual posso "referir" a aquilo que não existe; não é esse o sentido usado aqui. Uma palavra mais antiga para aquilo que chamo "representação" ou "referência" é *denotação*.
Em segundo lugar, sigo o costume dos lógicos contemporâneos e uso "existe" com o significado de "existe no passado, no presente ou no futuro". Assim, Winston Churchill "existe", e podemos nos "referir a" ou "representar" Winston Churchill

qualquer coisa que seja. Do mesmo modo (no que diz respeito a várias coisas complicadas) as características de Winston Churchill não são suficientes para fazer com que algo represente ou refira Winston Churchill. E nem é necessário: na nossa comunidade, a forma impressa "Winston Churchill", as palavras proferidas "Winston Churchill", e muitas outras coisas são usadas para representar Churchill (ainda que de maneira não pictórica), embora não tenham o tipo de similaridade com Churchill – mesmo a linha traçada – que um retrato tem. Se a *similaridade* não é nem necessária nem suficiente para fazer com que algo represente outra coisa, como *alguma coisa* poderia ser necessária ou suficiente a esse propósito? De que modo, afinal de contas, uma coisa pode representar (ou "corresponder a", etc.) outra coisa diferente?

A reposta parece fácil. Suponha que a formiga tivesse visto Winston Churchill, e suponha que ela tivesse inteligência e habilidade para desenhar um retrato dele. Suponha que ela fez a caricatura *intencionalmente*. A linha, nesse caso, representaria Churchill.

Por outro lado, suponha que a linha tivesse a forma WINSTON CHURCHILL. E suponha que isso tenha sido apenas um acidente (ignorando a improbabilidade envolvida). Assim, a "forma impressa" WINSTON CHURCHILL *não* teria representado Churchill, muito embora a forma impressa hoje represente Churchill quase que em todos os livros.

Portanto, pode parecer que o que é necessário para a representação, ou o que é principalmente necessário para a representação, é a *intenção*.

Mas para que a intenção de que *algo*, ainda que de uma linguagem privada (em que as palavras "Winston Churchill" fossem ditas em minha mente e não em voz alta), *representasse* Churchill, eu teria de ter sido capaz de *pensar sobre* Churchill em primeiro lugar. Se linhas na areia, barulhos, etc., não puderem "por si mesmos" representar algo, então de que modo é que essas formas pensadas conseguem "em sim mesmas" representar algo, se é que conseguem? De que modo o pensamento atinge e "apreende" aquilo que é externo?

No passado, alguns filósofos saltaram desse tipo de consideração àquilo que consideraram uma prova de que a mente é, *em natureza, essencialmente não física*. O argumento é simples: aquilo que dissemos sobre a linha traçada pela formiga se aplica a qualquer objeto físico. Nenhum objeto físico pode, em si mesmo, referir a uma coisa ao invés de outra; não obstante, os *pensamentos em nossa mente* obviamente referem-se, de maneira bem sucedida, a uma coisa ao invés de outra. Assim, os pensamentos (e, por conseguinte, a mente) são de uma natureza essencialmente diferente da dos objetos físicos. Os pensamentos têm a característica da *intencionalidade* – eles podem se referir a coisas; nada físico tem "intencionalidade", salvo quando a intencionalidade deriva de algum emprego dessa coisa física pela mente. Ou, pelo menos, se diz ser assim. Mas

muito embora há já não esteja vivo.

isso é apressado; postular poderes misteriosos da mente nada resolve. O problema, contudo, é bastante real. Como são possíveis a intencionalidade e a referência?

As teorias mágicas da referência

Vimos que o "retrato" feito pela formiga não tem uma conexão necessária com Winston Churchill. O mero fato de esse "retrato" possuir uma "semelhança" com Winston Churchill não o torna uma imagem real, nem mesmo uma representação de Churchill. A menos que a formiga fosse inteligente (o que não é) e tivesse conhecimento de Churchill (o que não tem), a linha por ela traçada não é uma imagem e nem mesmo uma representação do que quer que seja. Alguns povos primitivos acreditam que algumas representações (*nomes*, em particular) têm uma conexão necessária com seus portadores; que saber o "verdadeiro nome" de alguém ou de algo lhe confere poder sobre eles. Esse poder advém da *conexão mágica* entre o nome e o portador do nome; mas uma vez que se perceba que um nome *apenas* tem uma conexão contextual, contingente e convencional com seu portador, fica difícil ver por que o conhecimento do nome deveria ter qualquer significado místico.

O que é importante perceber é que aquilo que vale para os retratos físicos também vale para as imagens mentais e também para as representações em geral; as representações mentais não têm uma conexão necessária com aquilo que representam, não mais do que as representação físicas o têm. A suposição contrária é uma volta ao pensamento mágico.

Talvez, no caso das *imagens* mentais, seja mais fácil de se compreender o que está em questão (talvez o primeiro filósofo a ter compreendido a enorme importância desse ponto, ainda que não o primeiro a efetivamente enfatizá-lo, foi Wittgenstein). Suponha que algures há um planeta onde os seres humanos tenham evoluído (ou sido deixados lá por alienígenas, ou a hipótese que mais lhe aprouver). Suponha que esses humanos, embora em muitos aspectos sejam como nós, nunca tenham visto *árvores*. Suponha que eles nunca tenham imaginado árvores (talvez a vida vegetal exista nesse planeta apenas em forma de bolor). Suponha que, um dia, um retrato de uma árvore caia acidentalmente de uma nave que por ali passava, sem qualquer contato com eles. Imagine-os intrigados com esse retrato. Que coisa no mundo é isso? Todo o tipo de especulação ocorre a eles: um edifício, um abrigo ou, até mesmo, alguma espécie de animal. Mas suponha que eles nunca tivessem se aproximado da verdade.

Para nós, o retrato é uma representação de uma árvore. Para esses humanos, o retrato representa apenas um estranho objeto de natureza e função desconhecidos. Suponha que

um deles, tendo visto o retrato, tenha uma imagem mental que seja exatamente como uma das minhas imagens mentais de uma árvore. A sua imagem mental não é um *representação de uma árvore*. É apenas uma representação de um objeto estranho (qualquer que seja) que o misterioso retrato representa.

Contudo, alguém poderia argumentar que a imagem mental seja, *de fato*, uma representação de uma árvore, pois, para início de conversa, o retrato que causou a imagem mental era, ele próprio, uma representação de uma árvore. Há uma cadeia causal das árvores reais à imagem mental, ainda que uma cadeia causal bastante estranha.

Poderíamos imaginar, contudo, a ausência dessa cadeia causal. Suponha que o "retrato da árvore" deixado pela nave espacial não era realmente um retrato de uma árvore, mas o resultado acidental de tinta derramada. Ainda que fosse exatamente como o retrato de uma árvore, não era, na verdade, um retrato de uma árvore, não mais do que a "caricatura" de Churchill feita pela formiga era um retrato de Churchill. Podemos mesmo imaginar que a espaçonave que deixou o "retrato" veio de um planeta que não conhecesse árvores. Assim, os humanos ainda teriam imagens mentais qualitativamente idênticas à minha imagem de uma árvore, mas que não seriam imagens que representassem árvores ou qualquer outra coisa mais.

O mesmo vale para as *palavras*. Um discurso no papel poderia parecer ser uma descrição perfeita de árvores, mas se tivesse sido produzido por macacos a bater aleatoriamente as teclas de uma máquina de escrever durante milhões de anos, então as palavras não se refeririam a qualquer coisa. Se houvesse uma pessoa que memorizasse essas palavras e as recitasse mentalmente sem entendê-las, então elas sequer se refeririam a algo quando pensadas na mente.

Imagine que a pessoa a dizer essas palavras tenha sido hipnotizada. Suponha que as palavras estejam em japonês, e que disseram a essa pessoa que ela entende japonês. Suponho que, ao pensar nessas palavras, ela tenha uma "sensação de compreensão" (muito embora, caso alguém interrompesse a sequência do seu pensamento e lhe perguntasse o que *significam* aquelas palavras, ela descobriria que não conseguiria responder). Talvez a ilusão fosse tão perfeita que a pessoa pudesse até tapear um telepata japonês! Mas se ela não conseguisse usar as palavras nos contextos corretos, responder às perguntas sobre o que ele estava "pensado", etc., então ela não as teria entendido.

Combinando esses contos de ficção científica que acabei de contar, podemos inventar um caso no qual alguém pensa palavras que de fato são uma descrição de árvores em alguma linguagem *e* que simultaneamente tem imagens mentais apropriadas, mas que *nem* entende

as palavras e nem sabe o que seja uma árvore. Podemos imaginar que as imagens mentais tivessem sido causadas por respingos de tinta (embora a pessoa tenha sido hipnotizada para pensar que fossem imagens de algo apropriado a seu pensamento – algo que não seria capaz de dizer o que é, caso lhe fosse perguntado). E podemos imaginar que a linguagem na qual a pessoa está pensando é uma linguagem que nem o hipnotizador nem o hipnotizado tivessem ouvido falar – talvez fosse apenas uma coincidência que essas "frases sem sentido", como supõe o hipnotizador, sejam uma descrição de árvores em japonês. Em suma, tudo o que se passa diante da mente da pessoa poderia ser qualitativamente idêntico ao que estivesse se passando na mente de um falante japonês que estivesse, *de fato*, pensando em árvores – mas não se referiria a árvores.

Tudo isso é realmente impossível, assim como é realmente impossível que macacos pudessem, por acaso, datilografar uma cópia de *Hamlet*. Isso significa que as probabilidades contra tal são tão altas, que nunca ocorrerá na realidade (pensamos nós). Mas não é logicamente impossível, e nem mesmo fisicamente impossível. *Poderia* acontecer (é compatível com as leis da física e, talvez, compatível com as condições efetivas do universo, caso houvesse seres inteligentes em outros planetas). E, se acontecesse, seria uma demonstração impressionante de uma verdade conceitual importante; que mesmo um sistema amplo e complexo de representações, tanto verbal quanto visual, não tem uma conexão mágica *intrínseca*, embutida, com aquilo que representa – uma conexão independente de como foi causado e de quais são as disposições do falante ou pensante. E isso é verdade, esteja o sistema de representações (palavras e imagens, no caso do exemplo) fisicamente realizado – as palavras estão escritas ou proferidas, e as imagens são imagens físicas –, esteja apenas realizado na mente. Palavras pensadas e retratos mentais não representam *intrinsecamente* aquilo a que se referem.

O caso dos cérebros numa cuba

Eis uma possibilidade de ficção científica que os filósofos discutem: imagine que um ser humano (você pode imaginar isso para si mesmo) tenha sido submetido a uma operação por um cientista maligno. O cérebro dessa pessoa (o seu cérebro) foi removido de seu corpo e colocado numa cuba com nutrientes que mantêm o cérebro vivo. As terminações nervosas foram conectadas a um supercomputador que causa na pessoa, cujo cérebro ela é, a ilusão de que tudo é perfeitamente normal. Parece haver pessoas, objetos, o céu, etc.; mas na verdade tudo o que as pessoas (você) estão experienciando é o resultado de impulsos elétricos

viajando do computador às terminações nervosas. O computador é tão engenhoso que se a pessoa tenta levantar a mão, um *feedback* logo o fará "ver" e "sentir" a mão sendo levantada. Ademais, o cientista maligno pode, variando o programa, fazer a vítima "experienciar" (ou alucinar) qualquer situação ou ambiente. Ele pode também obliterar a memória da operação cerebral, de modo que a vítima veja a si próprio com sempre tendo estado naquele ambiente. Pode até parecer à vítima que ela está sentada e lendo estas palavras sobre essa divertida, embora absurda, suposição de que há um cientista maligno que remove o cérebro das pessoas de seus corpos e os coloca numa cuba com nutrientes que mantém o cérebro vivo. As terminações nervosas, supostamente, estão conectadas num supercomputador que causa na pessoa, cujo cérebro ela é, a ilusão de que...

Esse tipo de possibilidade, quando mencionada numa aula de Teoria do Conhecimento, tem certamente o propósito de levantar o problema clássico do ceticismo sobre o mundo externo de maneira contemporânea (*como você sabe que não está nessa situação*?). Mas essa situação é também um artifício útil para se levantar questões sobre a relação mente/mundo.

Ao invés de ter apenas um cérebro numa cuba, poderíamos imaginar que todos os seres humanos (talvez todos os seres sencientes) são cérebros numa cuba (ou sistemas nervosos numa cuba, no caso de alguns seres com apenas um sistema nervoso mínimo contarem como "sencientes"). É claro que o cientista maligno teria de estar do lado de fora – ou onde estaria? Talvez não haja qualquer cientista maligno, talvez (embora absurdo) seja apenas o caso de o universo consistir de uma maquinaria automática que tende a uma cuba cheia de cérebros e sistemas nervosos.

Isso, por sua vez, nos leva a supor que a maquinaria automática foi programada para nos causar uma alucinação *coletiva*, ao invés de nos causar, separadamente, diversas alucinações não relacionadas. Assim, quando me parece que estou conversando com você, parece-te que você está ouvindo as minhas palavras. É claro que não é o caso que as minhas palavras chegam de fato aos seus ouvidos – pois você não tem ouvidos (reais), e nem eu tenho uma boca e uma língua reais. Antes, quando produzo minhas palavras, o que acontece é que os impulsos eferentes vão do meu cérebro até o computador, que faz com que eu "ouça" minha própria voz proferindo aquelas palavras e "sinta" a língua se movimentar, etc., e também faz com que você "ouça" as minhas palavras, "veja"-me falando, etc. Nesse caso, estamos, em certo sentido, nos comunicando de fato. Não estou errado sobre a sua existência real (apenas sobre a existência de seu corpo e do "mundo externo", à parte os cérebros). De certo ponto de vista, não importa sequer que "todo o mundo" seja uma alucinação coletiva; pois, afinal de contas, você, de fato, houve as minhas palavras quando as digo a você, ainda que o mecanismo não seja aquele que supúnhamos ser (certamente, se estivéssemos a fazer

amor, e não apenas estivéssemos conversando, então a sugestão de que somos apenas dois cérebros numa cuba seria perturbadora).

Gostaria, agora, de fazer uma pergunta que parecerá um tanto tola e óbvia (pelo menos a alguns, incluindo alguns filósofos sofisticados), mas que rapidamente nos levará às profundezas filosóficas. Suponha que toda essa estória seja de fato verdadeira. Poderíamos, caso fossemos cérebros numa cuba, *dizer* ou *pensar* que o somos?

Argumentarei que a resposta é "Não, não poderíamos". Na verdade, argumentarei que a suposição de que somos, de fato, cérebros numa cuba, embora não viole qualquer lei da física e seja perfeitamente consistente com tudo o que experienciamos, não pode ser verdadeira. *Ela não pode ser verdadeira* pois é, de certo modo, autorrefutante.

O argumento que irei apresentar não é usual, e levei vários anos a me convencer de que ele estivesse realmente correto. Mas é um argumento correto. O que o torna tão estranho é que ele esteja conectado com algumas das mais profundas questões filosóficas (tal argumento me ocorreu pela primeira vez quando eu estava pensando sobre um teorema da lógica moderna, o "Teorema de Skolem-Löwenheim", e, repentinamente, vi uma conexão entre esse teorema e alguns argumentos nas *Investigações Filosóficas* de Wittgenstein).

Uma "suposição autorrefutante" é aquela cuja verdade implica a sua própria falsidade. Por exemplo, considere a tese de que *todas as afirmações gerais são falsas*. Essa é uma afirmação geral. Portanto, se for verdadeira, tem de ser falsa. Por conseguinte, é falsa. Às vezes chama-se a uma tese "autorrefutante" se *a sua suposição, assentida ou enunciada*, implica a sua falsidade. Por exemplo, "Eu não existo" é autorrefutante se pensado por mim (para qualquer "mim"). Assim, alguém que pensa sobre si pode estar certo de que existe (como argumentou Descartes).

O que vou mostrar é que a suposição de que somos cérebros numa cuba tem exatamente essa propriedade. Se a pudermos considerar como verdadeira ou falsa, então ela não é verdadeira (mostrarei). Assim, ela não é verdadeira.

Antes de oferecer o argumento, consideremos por que parece tão estranho que tal argumento possa ser oferecido (pelo menos aos filósofos que subscrevem a concepção de verdade como "cópia"). Concedemos que é compatível com as leis da física que houvesse um mundo no qual todos os seres sencientes são cérebros numa cuba. Como dizem os filósofos, há um "mundo possível" no qual todos os seres sencientes são cérebros numa cuba (essa conversa sobre "mundo possível" soa como se houvesse um *lugar* onde qualquer proposição absurda fosse verdadeira, e é esse o motivo pelo qual isso possa ser tão enganador na filosofia). Os humanos, nesse mundo possível, têm exatamente a mesma experiência

que *nós* temos. Eles têm o mesmo pensamento que nós (pelo menos as mesmas palavras, imagens, formas mentais, etc. passam pelas suas mentes). Contudo, estou afirmando que há um argumento que possa ser oferecido que mostre que não somos cérebros numa cuba. Como? E por que não poderiam as pessoas do mundo possível que *são*, de fato, cérebros numa cuba oferecê-lo também?

A reposta será (basicamente) esta: muito embora as pessoas nesse mundo possível possam pensar e "dizer" quaisquer palavras que possamos pensar e dizer, elas não podem (defenderei) se *referir* ao que nós podemos. Em particular, elas não podem pensar ou dizer que são cérebros numa cuba (*mesmo ao pensar "nós somos cérebros numa cuba"*).

O teste de Turing

Suponha que alguém invente um computador que pode, de fato, ter uma conversa com alguém (com tantos sujeitos quanto uma pessoa inteligente poderia). Como se poderia decidir se tal computador é "consciente"?

O lógico britânico Alan Turing propôs o seguinte teste:[3] faça com que alguém tenha uma conversa com o computador e com uma pessoa que ela não conheça. Se ela não conseguir dizer qual é o computador e qual é o humano, então (suponha que o teste possa ser repetido uma quantidade suficiente de vezes com diferentes interlocutores) o computador é consciente. Em suma, uma máquina de computar é consciente se puder passar no "Teste de Turing" (as conversas não são cara a cara, é claro, uma vez que o interlocutor não pode saber a aparência visual de ambos os seus parceiros. E nem se pode usar a voz, uma vez que a voz mecânica poderia soar diferente de uma voz humana. Imagine que o interlocutor digita as suas afirmações, perguntas, etc., e os dois parceiros – a máquina e a pessoa – respondem via um teclado eletrônico. Além disso, a máquina pode *mentir* – perguntada "Você é uma máquina", ela poderia responder, "Não, sou um assistente daqui do laboratório").

A ideia de que esse teste seja de fato um teste definitivo de consciência tem sido criticada por diversos autores (que não são por isso, a princípio, hostis à ideia de que uma máquina pudesse ser consciente). Mas esse não é o nosso tópico aqui. Gostaria de usar a ideia geral do teste de Turing, a ideia geral de um *teste dialógico de competência*, para um propósito diferente, o propósito de explorar a noção de *referência*.

3 - A. M. Turing, "Computing Machinery and Intelligence", *Mind* (1950), reimpresso em A. R. Anderson (ed.), *Minds and Machines*.

Imagine uma situação na qual o problema não é o de determinar se o parceiro é de fato uma pessoa ou uma máquina, mas antes determinar se o parceiro usa as palavras para referir assim como nós. O teste óbvio é, novamente, começar uma conversa, e, caso nenhum problema surja, se o parceiro "passa" no sentido de ser indistinguível de alguém que fale certificadamente a mesma língua de maneira avançada, refere-se aos tipos usuais de objetos, etc., para concluir que o parceiro refere a objetos assim como nós. Quando o propósito do teste de Turing for como o descrito acima, isto é, determinar a existência de referência (compartilhada), vou me referir ao teste como *Teste de Turing para a Referência*. E, assim como os filósofos têm discutido a questão de se o teste de Turing original é um teste *definitivo* para a consciência, *i.e.* a questão de se uma máquina que "passa" no teste não apenas uma vez, mas regularmente, é *necessariamente* consciente, da mesma maneira quero discutir a questão se o Teste de Turing para a Referência anteriormente sugerido é um teste definitivo para a referência compartilhada.

A resposta será "Não". O Teste de Turing para a Referência não é definitivo. Certamente que é um excelente teste na prática; mas não é logicamente impossível (embora seja altamente improvável) que alguém que pudesse passar no Teste de Turing para a Referência e não se referisse a qualquer coisa. Segue-se disso, como veremos, que podemos estender a nossa observação de que as palavras (e textos e discursos completos) não têm uma conexão necessária com os seus referentes. Ainda que consideremos não as palavras mas as regras que decidem quais palavras podem apropriadamente ser produzidas em certos contextos – ainda que consideremos, no jargão da computação, *programas para usar palavras* – a menos que os próprios programas *se refiram a algo extralinguístico*, não haverá, contudo, referência determinada que essas palavras possuam. Esse será um passo crucial no processo de atingir a conclusão de que os Habitantes do Mundo do Cérebro numa Cuba não podem se referir a qualquer coisa externa (e, por conseguinte, não podem dizer *que* são Habitantes do Mundo do Cérebro numa Cuba).

Suponha, por exemplo, que estou numa situação de Turing (jogando o "Jogo da Imitação", na terminologia de Turing) e meu parceiro seja, de fato, uma máquina. Suponha que essa máquina seja programada para vencer o jogo ("passe" no teste). Imagine que a máquina foi programada para dar belas respostas em português às afirmações, perguntas, observações, etc., mas que não tem órgãos sensoriais (nada além das conexões com o meu teclado eletrônico) e nem órgãos motores (nada além do teclado eletrônico) (tanto quanto posso entender, Turing não supõe que a possessão de órgãos dos sentidos ou de órgão motores seja necessário para consciência ou inteligência). Suponha que não apenas a máquina careça de olhos e ouvidos eletrônicos, etc., mas que também não há recursos no programa da máquina, o programa para jogar o Jogo da Imitação, para incorporar *inputs* de

tais órgãos, ou para controlar um corpo. O que dizer de tal máquina?

A mim parece evidente que não podemos e não deveríamos atribuir referência a tal dispositivo. É verdade que a máquina pode discursar belamente sobre, digamos, a vista na Nova Inglaterra. Mas não conseguiria reconhecer uma macieira ou uma maçã, uma montanha ou uma vaca, um campo ou um campanário, caso estivesse na frente de algum deles.

O que temos é um dispositivo para produzir frases em respostas a outras frases. Mas nenhuma dessas frases está de todo conectada ao mundo real. *Se ligássemos duas dessas máquinas e as deixássemos jogar o Jogo da Imitação uma com a outra, elas continuariam a "enrolar" uma a outra para sempre, ainda que o resto do mundo desaparecesse*! Não há mais razão para considerar a conversa da máquina sobre maçãs como se referindo a maçãs do mundo real do que há para considerar os "traços" da formiga com se referindo a Winston Churchill.

O que produz aqui a ilusão de referência, significado, inteligência, etc., é o fato de que há uma convenção de representação da qual *nós* aceitamos de que o discurso da máquina se refere a maçãs, campanários, a Nova Inglaterra, etc. Similarmente, há a *ilusão* de que a formiga fez uma caricatura de Churchill pela mesma razão. Mas nós somos capazes de perceber, manipular, lidar com maçãs e campos. A nossa conversa sobre maçãs e campos está intimamente conectada com as nossas transições *não-verbais* com as maçãs e os campos. Há "regras de entrada da linguagem" que nos levam de experiências de maçãs a elocuções como "Vejo uma maçã", e "regras de saída da linguagem" que nos levam de decisões expressas na forma linguística ("Estou indo comprar maçãs") a outras ações além da fala. Na ausência de regras de entrada ou saída da linguagem, não há razão para considerar a conversa da máquina (ou das duas máquinas, no caso em que pensamos nas duas máquinas jogando do Jogo da Imitação uma com outra) algo mais do que um jogo sintático. Um jogo sintático que *se assemelha* ao discurso inteligente, na verdade; mas apenas o tanto quanto (e não mais que isso) os traços feitos pela formiga se assemelham a uma sarcástica caricatura.

No caso da formiga, poderíamos ter argumentado que ela teria traçado a mesma linha, ainda que Winston Churchill nunca tivesse existido. No caso da máquina, não podemos usar exatamente o mesmo argumento; se as maçãs, as árvores, os campanários e campos não tivessem existido, então, presumivelmente, os programadores não teriam produzido esse mesmo programa. Embora a máquina não *perceba* maçãs, campos ou campanários, os seus criadores perceberam. Há uma conexão causal entre a máquina e as maçãs do mundo real, etc., por meio da experiência e conhecimento perceptual dos criadores. Mas tal conexão fraca dificilmente pode ser suficiente para a referência. Não apenas é logicamente possível, embora fantasticamente improvável, que a mesma máquina *pudesse* ter existido ainda que as maçãs, os campos e os campanários não; mais importante, a máquina é completamente

insensível à existência continuada das maçãs, dos campos, dos campanários, etc. Ainda que todas essas coisas *deixassem* de existir, a máquina ainda discursaria alegremente do mesmo jeito. É por isso que a máquina não pode ser considerada com se referindo a algo.

O ponto relevante para a nossa discussão é o de que nada há no Teste de Turing que exclua uma máquina programada para fazer nada *mais* do que jogar o Jogo da Imitação, e que exclua que tal máquina *claramente* não se refira a qualquer coisa, não mais do que um gravador se refere.

Cérebros numa cuba (novamente)

Comparemos os hipotéticos "cérebros numa cuba" com as máquinas que acabamos de descrever. Há, obviamente, diferenças importantes. Os cérebros numa cuba não têm órgãos sensoriais, mas têm *suporte* para tais órgãos; isto é, há terminações nervosas aferentes, há *inputs* dessas terminações nervosas aferentes, e esses *inputs* figuram no "programa" dos cérebros na cuba assim como figuram no programa dos nossos cérebros. Os cérebros numa cuba são *cérebros*; ademais, eles são cérebros *em funcionamento*, e funcionam por meio das mesmas regras que os cérebros do mundo efetivo funcionam. Por essas razões, pareceria absurdo negar consciência ou inteligência a eles. Mas o fato de que são conscientes e inteligentes não quer dizer que as suas palavras se refiram àquilo que as nossas palavras se referem. A questão que nos interessa é esta: as suas verbalizações contendo, digamos, a palavra "árvore" se referem realmente a *árvores*? De modo mais geral: eles podem se referir a objetos *externos* (em oposição, por exemplo, aos objetos na imagem produzida pela maquinaria automática)?

Para ajustar as nossas ideias, especifiquemos que a maquinaria automática veio a existir supostamente de algum tipo de acaso ou coincidência cósmica (ou, talvez, tenha sempre existido). Nesse mundo hipotético, a própria maquinaria automática supostamente não tem criadores inteligentes. Na verdade, como dissemos no início, podemos imaginar que todos os seres sencientes (ainda que minimamente sencientes) estão dentro da cuba.

Essa suposição não ajuda. Pois não há conexão entre a *palavra* "árvore" como usada por esses cérebros e as árvores reais. Eles continuariam a usar a palavra "árvore" como usam, teriam os pensamentos que têm, teriam as imagens que têm, ainda que não houvesse árvores reais. As suas imagens, palavras, etc. são qualitativamente idênticas às imagens, palavras, etc. que representam árvores no *nosso* mundo; mas já vimos (novamente a formiga!) que similaridade qualitativa a algo que representa um objeto (Winston Churchill ou uma

árvore) não faz de algo uma representação por si próprio. Em suma, os cérebros numa cuba não estão pensando sobre árvores reais quando pensam "há uma árvore à minha frente", porque nada há em virtude do qual o seu pensamento "árvore" represente árvores reais.

Caso isso pareça apressado, reflita sobre o seguinte: vimos que as palavras não necessariamente se referem a árvores, ainda que estejam dispostas numa sequência que seja idêntica a um discurso que (estivesse a ocorrer em uma de nossas mentes) fosse inquestionavelmente *sobre árvores* no mundo efetivo. Nem o "programa", no sentido das regras, práticas, disposições dos cérebros ao comportamento verbal, refere-se necessariamente a árvores ou produz a referência a árvores por meio das conexões que estabelece entre palavras e palavras, ou sugestões *linguísticas* e respostas *linguísticas*. Se esses cérebros pensam sobre, referem-se a, representam árvores (árvores reais, fora da cuba), então isso tem de ser por causa do modo como o "programa" conecta o sistema da linguagem a *inputs* e *outputs* não-verbais. Há, de fato, tais *inputs* e *outputs* não-verbais no Mundo do Cérebro numa Cuba (as terminações nervosas eferentes e aferentes novamente!), mas também vimos que os "dados dos sentidos" produzidos pela maquinaria automática não representam árvores (ou algo externo) mesmo quando eles se assemelham exatamente às nossas imagens de árvores. Assim como um salpico de tinta poderia se assemelhar a um retrato de uma árvore sem *ser* um retrato de uma árvore, do mesmo modo, vimos, um "dado do sentido" poderia ser qualitativamente idêntico a uma "imagem de uma árvore" sem ser uma imagem de uma árvore. Como é que pode o fato de, no caso dos cérebros numa cuba, a linguagem conectada pelo programa com os *inputs* sensoriais que não representam intrínseca ou extrinsecamente árvores (ou algo externo) faça com que todo o sistema de representações, a linguagem em uso, refira-se a ou represente árvores ou algo externo?

A resposta é que não pode. Todo o sistema de dados dos sentidos, de sinais motores às terminações eferentes, e de pensamento verbalmente ou conceitualmente mediado conectado pelas "regras de entrada da linguagem" aos dados dos sentidos (ou seja lá o que for) como *inputs* e pelas "regras de saída da linguagem" aos sinais motores como *outputs*, não tem mais conexão com *árvores* do que as linhas feitas pela formiga tem com Winston Churchill. Uma vez que percebemos que a *similaridade qualitativa* (denotando, caso queiras, identidade qualitativa) entre os pensamentos dos cérebros numa cuba e os pensamentos de alguém no mundo efetivo já não implica o compartilhamento de referência, não é difícil ver que não há bases de todo em todo para se considerar o cérebro numa cuba como se referindo a coisas externas.

A premissa do argumento

Ofereci o argumento prometido para mostrar que os cérebros numa cuba não podem pensar ou dizer que são cérebros numa cuba. Falta apenas torná-lo explícito e examinar a sua estrutura.

Por aquilo que foi dito, quando o cérebro numa cuba (no mundo onde todo ser senciente é e sempre foi uma cuba) pensa "Há uma árvore à minha frente", o seu pensamento não se refere às árvores reais. De acordo com algumas teorias que discutiremos, poderia se referir às árvores na imagem, ou às características do programa que são responsáveis pelos impulsos elétricos. Essas teorias não são excluídas pelo que foi dito, pois há uma conexão causal próxima entre o uso da palavra "árvore" no português da cuba e a presença das árvores na imagem, a presença dos impulsos elétricos de certo tipo, e a presença de ceras características no programa da máquina. De acordo com essas teorias, o cérebro está *correto*, e não *errado* em pensar "Há uma árvore à minha frente". Dado ao que "árvore" se refere no português da cuba e ao que "em frente de" se refere, supondo que uma dessas teorias esteja correta, então as condições de verdade para "Há uma árvore à minha frente" quando ocorre no português da cuba são simplesmente as de que uma árvore na imagem esteja "em frente a" o "mim" em questão – na imagem – ou, talvez, que o tipo de impulso elétrico que normalmente produz essa experiência esteja vindo da maquinaria automática, ou, talvez, que a característica da maquinaria que supostamente produz a experiência da "árvore em minha frente" esteja operando. E essas condições de verdade são certamente cumpridas.

Pelo mesmo argumento, "cuba" se refere, no português da cuba, a cubas na imagem, ou a algo de algum modo relacionado (impulsos elétricos ou características de programa), mas não certamente a cubas reais, uma vez que o uso de "cuba" no português da cuba não tem conexão causal com as cubas reais (com exceção da conexão em que os cérebros numa cuba não seriam capazes de usar a palavra "cuba", caso não fosse pela presença de uma cuba particular – a cuba em que eles estão; mas essa conexão é obtida entre o uso de *cada* palavra no português da cuba e essa cuba particular; não é uma conexão especial entre o uso da palavra *particular* "cuba" e as cubas). Similarmente, "fluído nutriente", no português da cuba, refere-se a um líquido, ou a algo relacionado (impulsos elétricos ou características do programa). Segue-se disso que se tal "mundo possível" for de fato o mundo efetivo, e nós formos de fato cérebros numa cuba, então aquilo que queremos dizer por "somos cérebros numa cuba" é que *somos cérebros numa cuba na imagem* ou algo do tipo (se é que dizemos algo). Mas parte da hipótese de que somos cérebros numa cuba é que não somos cérebros numa cuba na imagem (*i.e.* aquilo que estamos a "alucinar" não é que somos cérebros numa cuba). Assim, se somos cérebros numa cuba, então a frase "Somos cérebros numa cuba" diz

algo falso (se é que o diz). Em suma, se somos cérebros numa cuba, então "Somos cérebros numa cuba" é falsa. Portanto, é (necessariamente) falsa.

A suposição de que tal possibilidade faz sentido surge da combinação de dois erros: (1) levar demasiadamente a sério a *possibilidade física*; e (2) usar inconscientemente uma teoria mágica da referência, uma teoria segundo a qual certas representações mentais referem-se necessariamente a coisas e tipos de coisas externos.

Há um "mundo fisicamente possível" no qual somos cérebros numa cuba – o que isso significa, exceto que há uma *descrição* de tal estado de coisas que é compatível com as leis da física? Assim como há uma tendência em nossa cultura (e tem sido assim desde o século dezessete) de considerar a *física* como a nossa metafísica, isto é, ver as ciências exatas como a tão estimada descrição da "verdadeira e última mobilha do universo", há, como uma consequência imediata, uma tendência a considerar a "possibilidade física" com a própria pedra de toque daquilo que poderia realmente ser de fato o caso. De acordo com tal perspectiva, a verdade é a verdade física; a possibilidade, a possibilidade física; e a necessidade, a necessidade física. Mas acabamos de ver, mesmo que apenas no caso de um exemplo bastante artificial, que tal perspectiva está errada. A existência de um "mundo fisicamente possível" no qual somos cérebro numa cuba (e sempre fomos e sempre seremos) não significa que realmente, efetivamente, pudéssemos *ser* cérebros numa cuba. O que exclui essa possibilidade não é a física, mas a *filosofia*.

Alguns filósofos, ávidos em asseverar e minimizar as pretensões de sua profissão (a típica mentalidade da filosofia anglo-saxônica no século vinte), diriam: "Certo. Você mostrou que algumas coisas que parecem ser possibilidades físicas são impossibilidades *conceituais*. O que há de tão surpreendente nisso?"

Ora, para ser sincero, o meu argumento pode ser descrito como um argumento "conceitual". Mas descrever a atividade filosófica como a busca por verdades "conceituais" faz tudo soar como se fosse uma *investigação sobre o significado das palavras*. E isso não é de todo o que estivemos a fazer.

Aquilo que estivemos a fazer foi considerar as *precondições* para o *pensamento acerca de algo*, para a *representação*, para a *referência*, etc. Investigamos essas precondições *não* por meio da investigação do significado dessas palavras e expressões (como um linguista faria, por exemplo), mas pelo *raciocínio a priori*. Não no velho sentido "absoluto" (uma vez que não afirmamos que as teorias mágicas da referência estão *a priori* erradas), mas no sentido de investigar aquilo que é *razoavelmente* possível *se aceitarmos* certas premissas gerais, ou fazer suposições teóricas bastante amplas. Tal procedimento não é nem "empírico" nem

completamente *a priori*, mas tem elementos de ambos os modos de investigação. A despeito da falibilidade do meu procedimento, e sua dependência de suposições que poderiam ser descritas como "empíricas" (*e.g.* a suposição de que a mente não tem acesso a propriedades ou coisas externas, a menos que fornecido pelos sentidos), ele tem uma relação íntima com aquilo que Kant chamou de investigação "transcendental"; pois é uma investigação, repito, das *precondições* da referência e, por conseguinte, do pensamento – precondições embutidas na natureza das nossas próprias mentes, embora não completamente (como Kant esperava) independente de suposições empíricas.

Uma das premissas do argumento é obvia: as teorias mágicas da referência estão erradas, e não apenas erradas para as representações físicas, mas também para as mentais. A outra premissa é que não podemos referir a certos tipos de coisas, *e.g. árvores*, se não tivermos qualquer interação causal com elas,[4] ou com as coisas em termos das quais elas podem ser descritas. Mas por que deveríamos aceitar essas premissas? Uma vez que constituem o enquadramento amplo no qual estou argumentando, é hora de examiná-las mais de perto.

As razões para rejeitar as conexões necessárias entre as representações e seus referentes

Mencionei anteriormente que alguns filósofos (o mais famoso deles, Brentano) atribuíram um poder à mente, a "intencionalidade", que é precisamente aquilo que nos permite *referir*. Evidentemente, rejeitei isso como solução. O que, contudo, me dá esse direito? Talvez eu não tenha sido apressado demais?

Esses filósofos não afirmam que podemos pensar sobre coisas ou propriedades externas sem usar representações de todo em todo. E o argumento que ofereci acima compara os dados dos sentidos visuais com o "retrato" feito pela formiga (o argumento no conto de ficção científica sobre o "retrato" de uma árvore, surgido de um respingo de tinha, e que produziu os dados dos sentidos qualitativamente similares às nossas "imagens visuais de árvores", mas que vêm desacompanhados de qualquer *conceito* de árvore) teria sido aceito como mostrando que essas *imagens* não necessariamente referem. Se há representações mentais que necessariamente referem (a coisas externas) elas têm de ser da mesma natureza dos *conceitos* e não da natureza de imagens. Mas o que são *conceitos*?

4 - Se os cérebros numa cuba terão conexão causal com árvores, digamos, *no futuro*, então, talvez, eles possam se referir agora a árvores por meio da descrição "as coisas às quais irei me referir como 'árvores' em tal e tal tempo futuro". Mas estamos a imaginar um caso no qual os cérebros numa cuba nuca saíram da cuba e, por isso, *nunca* tiveram conexão causal com árvores, etc.

Ao introspectarmos, não percebemos "conceitos" fluindo em nossas mentes como tais. Interrompa o fluxo do pensamento quando ou onde quiser e tudo o que teremos serão palavras, imagens, sensações, sentimentos. Quando penso em voz alta, não penso os meus pensamentos duas vezes. Escuto as minhas palavras assim como você. Para dizer a verdade, é diferente para mim quando profiro palavras que acredito e quando profiro palavras que não acredito (mas, às vezes, quando estou nervoso, ou frente a uma audiência hostil, sinto como se eu estivesse a mentir quando sei que estou dizendo a verdade); e é diferente quando profiro palavras que entendo e quando profiro palavras que não entendo. Mas posso imaginar sem dificuldade alguém a pensar exatamente essas palavras (no sentido de dizê-las em sua mente) e a ter exatamente a sensação de entendimento, de asserção, etc. que tenho, mas que perceba, logo em seguida (ou ao ser despertado por um hipnotizador), que não entendeu aquilo que acabou de passar em sua mente, que nem sequer entendeu a língua na qual essas palavras estão. Não digo que isso seja muito provável; digo simplesmente que nada há de ininteligível com isso. E o que isso mostra não é que conceitos *são* palavras (ou imagens, ou sensações, etc.), mas que atribuir um "conceito" ou um "pensamento" a alguém é completamente diferente de atribuir uma "apresentação" mental, alguma entidade ou evento introspectível, a essa pessoa. Conceitos não são apresentações mentais que intrinsecamente referem a objetos externos pela mesma razão decisiva que não são sequer representações mentais. Conceitos são signos usados de certas formas; os signos podem ser públicos ou privados, entidades mentais ou físicas, mas mesmo quando são "mentais" e "privados", o próprio signo, à parte de seu uso, não é um conceito. E signos em si mesmos não referem intrinsecamente.

Podemos ver isso por meio de um experimento mental bastante simples. Suponha que você é como a mim mesmo e não consegue distinguir um olmo de uma faia. Dizemos, contudo, que a referência de "olmo" em minha fala é a mesma referência de "olmo" na fala do restante das pessoas, *viz*. olmos, e que o conjunto de todas as faias é a extensão de "faia" (*i.e.* o conjunto das coisas dais quais "faia" é verdadeiramente predicada) tanto na sua fala quanto na minha. É realmente crível que a diferença entre aquilo a que "olmo" se refere e aquilo a que "faia" se refere seja produzida por uma diferença nos nossos *conceitos*? O meu conceito de olmo é exatamente o mesmo que meu conceito de faia (tenho de confessar) (isso mostra que a determinação da referência é social e não individual, a propósito; você e eu diferimos dos especialistas que *podem* distinguir olmos de faias). Se alguém heroicamente tentar manter que a diferença entre as referências de "olmo" e "faia" na *minha* fala é explicada pela diferença em meu estado psicológico, então façamos com que ele imagine uma Terra Gêmea onde as palavras são trocadas. A Terra Gêmea é muito parecida com a Terra; na verdade, com a exceção de que "olmo" e "faia" são intercambiáveis, o leitor pode supor que

a Terra Gêmea é exatamente como a Terra. Suponha que eu tenha um *Doppelganger* na Terra Gêmea que é, molécula a molécula, idêntico a mim (no sentido em que duas gravatas podem ser "iguais"). Se você é um dualista, então suponha que o meu *Doppelganger* tem os mesmos pensamentos verbalizados que eu, que tem os mesmo dados dos sentidos, as mesmas disposições, etc. É absurdo pensar que o seu estado psicológico seja um pouco diferente do meu: contudo, a sua palavra "olmo" representa *faias*, e a minha palavra "olmo" representa olmos. (Similarmente, se a "água" na Terra Gêmea for um líquido diferente – digamos, XYZ e não H_2O – então "água" representa um líquido diferente quando usada na Terra Gêmea e quando usada na Terra, etc.). Contrário a uma doutrina que tem nos acompanhado desde o século dezessete, *os significados simplesmente não estão na cabeça*.

Vimos que possuir um conceito não é uma questão de possuir imagens (digamos, de árvores – ou mesmo imagens, "visuais" ou "acústicas", de frases, ou de discursos completos), uma vez que podemos possuir qualquer sistema de imagens e não possuir a *habilidade* de usar as frases em situações apropriadas (considere tanto fatores linguísticos – aquilo que foi dito antes – quanto fatores não linguísticos como aquilo que determina o que é "situacionalmente apropriado"). Um homem pode ter todas as imagens que você quiser e, ainda assim, ficar completamente perdido se alguém lhe disser "aponte-me uma árvore", ainda que haja muitas árvores ao redor. Ele pode até ter a imagem daquilo que tem de fazer e, ainda assim, não saber o que tem de fazer. Pois a imagem, se não for acompanhada pela habilidade de agir de certo modo, é apenas uma *imagem*, e agir de acordo com uma imagem é em si uma habilidade que alguém pode ter ou não. (O homem poderia imaginar-se a apontar para uma árvore, mas apenas a fim de contemplar algo logicamente possível; ele próprio apontando para a árvore depois que alguém produziu a sequência de sons – para ele destituída de sentido – "aponte-me uma árvore, por favor."). Ele ainda não saberia que tinha de apontar para uma árvore, e ainda não *entenderia* "aponte-me uma árvore".

Considerei a habilidade de usar certas frases como sendo o critério para a posse de um conceito bem desenvolvido, mas isso pode ser facilmente liberalizado. Poderíamos permitir que o simbolismo consistisse de elementos que não são palavras na linguagem natural, por exemplo, e poderíamos permitir tais fenômenos mentais como imagens e outros tipos de eventos internos. O que é essencial é que tais coisas tivessem a mesma complexidade, a habilidade de serem combinados uns com os outros, etc., que as palavras de uma linguagem natural têm. Pois, embora uma apresentação particular – de um *flash* azul, digamos – pudesse servir a um matemático particular como uma expressão interna de uma prova completa do Teorema dos Números Primos, ainda não haveria a tentação de dizer isso (e seria falso dizê-lo), caso o matemático não pudesse apresentar o seu "flash azul" em passos separados e conexões lógicas. Mas não importa que tipo de fenômenos internos

aceitamos como possíveis *expressões* de pensamento, argumentos exatamente similares ao anterior mostrarão que não são os fenômenos em si mesmos que constituem o significado, mas, antes, a habilidade do sujeito em *empregar* esses fenômenos, de produzir o fenômeno correto nas circunstancias corretas.

Isso é uma versão bastante abreviada do argumento de Wittgenstein nas *Investigações Filosóficas*. Se estiver correto, então a tentativa de entender o pensamento por meio daquilo a que se chama investigação "fenomenológica" está fundamentalmente errado; pois o que os fenomenólogos não conseguem ver é que aquilo que eles descrevem é a *expressão* interna do pensamento, mas que o *entendimento* da expressão – o entendimento que alguém tem de seus próprios pensamentos – não é uma *ocorrência*, mas, sim, uma *habilidade*. O nosso exemplo de um homem fingindo pensar em japonês (e enganando o telepata japonês) já mostrou a futilidade de uma abordagem fenomenológica ao problema do *entendimento*. Pois, ainda que haja alguma qualidade introspectível que se apresente quando, e apenas quando, alguém *de fato* entenda (isso parece falso na introspecção, na verdade), ainda assim essa qualidade *se relaciona* apenas ao entendimento, e é ainda possível que o homem tapeando o japonês tenha também essa qualidade e, contudo, não entenda uma palavras de japonês.

Por outro lado, considere a hipótese, perfeitamente possível, de um homem que não tem qualquer "monólogo interior". Ele fala um inglês perfeitamente bom, e, caso peçam a sua opinião sobre um dado assunto, ele a emitirá precisamente. Mas ele nunca pensa (em palavras, imagens, etc.) quando não está falando em voz alta; nem algo "passa pela sua cabeça", exceto (claro) que ele ouve a própria voz ao falar, e tem as impressões comuns de seus arredores, e mais uma "sensação de entendimento" geral (talvez ele tenha o hábito de falar consigo próprio). Quando ele escreve uma carta ou vai ao mercado, etc., ele não está tendo um "fluxo de pensamento" interno; mas as suas ações são inteligentes e dotadas de propósito, e se alguém vai até ele e pergunta "O que você está fazendo?" ele dará uma resposta perfeitamente coerente.

Esse homem parece perfeitamente imaginável. Ninguém hesitaria em dizer que ele fosse consciente, que não curtisse rock'n'roll (caso expressasse com frequência uma forte aversão ao rock'n'roll, etc., só porque ele não tem pensamentos conscientes, exceto quando fala em voz alta.

O que se segue disso é que (a) nenhum conjunto de eventos mentais – imagens ou acontecimentos e qualidades mentais mais "abstratos" – *constitui* o entendimento; e (b) nenhum conjunto de eventos mentais é *necessário* para o entendimento. Em particular, *conceitos não podem ser idênticos a objetos mentais de qualquer tipo*. Pois, supondo que por um objeto mental estamos designando algo introspectível, já vimos que, seja ele o que for, pode estar ausente num homem que entende a palavras apropriadas (e, por isso, tem o conceito

bem desenvolvido), e presente num homem que não tem conceito de todo em todo.

Voltando agora à nossa crítica das teorias mágicas da referência (um tópico que também interessava a Wittgenstein), vimos que, por um lado, aqueles "objetos mentais" que *podemos* detectar introspectivamente – palavras, imagens, sensações, etc. – não referem intrinsecamente, não mais do que o retrato feito pela formiga (e pelas mesmas razões), ao passo que as tentativas de postular objetos mentais, "conceitos", que têm uma conexão necessária com os seus referentes, e que apenas fenomenólogos treinados podem detectar, cometem uma asneira *lógica*; pois conceitos são (pelo menos em parte) *habilidades*, e não ocorrências. A doutrina de que há apresentações mentais que necessariamente referem a coisas externas não é apenas má ciência natural; é também má fenomenologia e confusão conceitual.