Análise da experiência conceptual do Quarto Chinês

Trabalho realizado por Pedro Fonseca,

Área C

Para o professor NUNO NABAIS

na cadeira de FILOSOFIA CONTEMPORÂNEA.

Apresentado em 20/7/98

Índice

NOTA: as notas em numeração romana encontram-se no final do documento

Num artigo de 1980 “Minds, Brains and Programs”,[1] Searle apresentou um argumento que levantou acesas discussões dentro e fora da área da filosofia cognitiva. Searle descreve o seu próprio trabalho na área da filosofia cognitiva da seguinte forma:

> “Eu refutei a ideia de que o cérebro é um computador digital e que a mente é um programa de computador num artigo … [em que] apresentei o agora bem conhecido Argumento do Quarto Chinês …. A força deste argumento tem sido extensamente debatida na literatura, e deve haver, pelo menos, algumas centenas de debates publicados sobre ela, mas nada nessa literatura me levou a supor que o Argumento do Quarto Chinês é algo menos do que uma refutação decisiva das pretensões mais ambiciosas da inteligência artificial.”[2]

O objectivo do nosso trabalho consiste em esclarecer em que medida e sobre que pressupostos podemos considerar que o Argumento do Quarto Chinês constitui uma ‘refutação decisiva’ e se ela de facto se aplica a alguma linha de investigação importante seguida hoje na área da IA (inteligência artificial) ou da filosofia cognitiva.

[1] In *Behavioral and Brain Sciences*, 1980, n.º 3.
[2]

1. O Argumento contra a IA forte.

O argumento de Searle divide-se em duas partes: a experiência conceptual propriamente dita do Quarto Chinês e as premissas e conclusões contra o que Searle apelida de 'IA forte' que essa experiência ilustra.

O argumento de Searle tem um objectivo que se pode enunciar de forma simples. Mostrar que os computadores podem simular mas não duplicar o pensamento humano.[3] Isto é, um computador, com um programa e um *hardware* correcto, pode vir a comportar-se *como se* fosse um ser humano (no limite, tornando-se indestinguível dele), no entanto, a esse comportamento aparentemente inteligente não pode corresponder qualquer *conteúdo mental*.

Por outro lado, este argumento não depende das limitações da tecnologia actual mas das limitações do computador enquanto tal. Isto é, nenhum computador pode ou poderá vir a ter estados mentais seja qual for a sua complexidade e rapidez e seja qual for o tamanho e complexidade do programa que utilize.[4]

O argumento de Searle tem, a princípio, três axiomas e uma conclusão.[5] Podemos expô-los da seguinte forma:

1. (A1) Os programas [de computador] são formais (sintácticos).

2. (A2) As mentes têm conteúdos mentais (semântica).

3. (A3) A sintaxe, por si só, não é constitutiva nem suficiente para a semântica

---

∴ (C1) Os programas não são constitutivos nem suficientes para as mentes.

A distinção e a relação entre sintaxe e semântica desempenham, como se vê, os papeis cruciais no argumento de Searle. A semântica é o significado dos símbolos, o

[3] V. *Mente, Cérebro e Ciência,* pp. 45-46.
[4] V. op.cit. p.45.
[5] Sigo a apresentação dada por Searle em "Is the Brain's Mind a Computer Program?", in *Scientific American* 1990 - 262: 26-31. Em *Mente, Cérebro e Ciência* não é tão claro que a conclusão apresentada aqui (C1) seja derivada apenas destas três premissas

facto de eles se referirem a qualquer coisa (o que Searle designa por intencionalidade[6]).

> "as séries [de símbolos] por si mesmas não têm qualquer significado. Se os meus pensamentos são *acerca de alguma coisa,* então as séries devem ter um *significado,* que faz com que os pensamentos sejam a propósito dessas coisas. Numa palavra, a mente tem mais do que uma sintaxe, possui também uma semântica."[7]

A sintaxe, por outro lado, corresponde a um conjunto de símbolos, com uma certa estrutura formal, mas sem significado. A ideia é que um conjunto articulado de símbolos, por si mesmo, não tem significado, isto é, 'não são acerca de qualquer coisa',[8] não apontam para nenhum objecto exterior a eles próprios. Por exemplo, num programa de computador "Os zeros e uns … são simples numerais; nem sequer estão em vez de números."[9]

Searle considera a primeira premissa como parte da definição de 'programa de computador', a segunda como um facto evidente e a terceira como uma verdade conceptual.[10] A conclusão seria, portanto, bastante sólida.

No entanto, das três premissas, só A2 parece ser trivial. Em relação a A1 não é imediatamente evidente que os zeros e uns (ou os impulsos eléctricos) não designem números ou não os possam designar perante o programa e *hardware* adequado.[i] Por outro lado A3 é contestada de um modo ainda mais generalizado. Ned Block por exemplo defende que pode haver intencionalidade sem consciência e vice-versa.[11] Para Block, os números dentro de um programa de computador designam de facto qualquer coisa dependendo do nível de descrição em que ocorram. Assim os numerais de um programa podem designar números[12] (ou palavras, noutro nível de descrição).[13]

---

[6] Cf. op.cit. p.21.
[7] Op.cit. p.39.
[8] Op.cit. p.38.
[9] Idem.
[10] Cf. Mente, Cérebro e Ciência, p. 48.
[11] Block é um dos principais críticos do funcionalismo, sobretudo desde o seu artigo de 1978 "Troubles with Functionalism", in *Perception and Cognition: Minnesota Studies in the Philosophy of Science*, do qual aparece um excerto em *Mind and Cognition,* Lycan, W. (ed), Blackwell, 1990, sob o título "«Qualia» –Based objections to Functionalism".
[12] Com excepção daqueles que operam ao nível dos 'processadores primitivos'. Neste caso devemos considerá-los como símbolos não interpretados. Uma discussão mais detalhada deste argumento de Block encontra-se no Apêndice 3. A posição de Block encontra-se em *The Mind as the Software of the Brain,* Caps. 1.3 a 3, esp. cap. 2.1 e penúltimo § do cap. 3.

E isto porque os zeros e uns começam por ser simplesmente sintácticos e adquirem uma semântica devido ao modo como são processados pelos processadores primitivos. Parece portanto que, para os críticos de Searle o verdadeiro alvo a abater é A3. De facto veremos que a principal crítica que se faz a Searle consiste em defender que o sistema como um todo pode ter propriedades diferentes do que cada uma das suas partes consideradas individualmente. Ou seja, seria possível, de um grande conjunto de operações sintácticas, organizadas de determinada maneira, atingir a semântica.

Vemos portanto que a afirmação de Searle de que A3 é uma verdade conceptual não é evidente para uma grande parte dos investigadores da ciência cognitiva. Searle vai demonstrar A3 através de uma experiência conceptual.[14] Entramos agora na experiência conceptual do Quarto Chinês.

[13] Uma discussão desta questão é proporcionada no Apêndice 1.

[14] A verdade de A1 normalmente não é posta em causa na literatura sobre este argumento. No entanto, na nossa leitura, a experiência conceptual de Searle é congruente quer com a falsidade quer com a verdade de A1 (ela não demonstra a verdade de A1), embora, no primeiro caso, possamos admitir a validade das pretensões da IA forte. Em parte, é isso que torna o exemplo de Searle tão intuitivamente convincente. Este assunto é discutido no Apêndice 2.

2. O exemplo do Quarto Chinês[15]

Imaginemos uma caixa pequena fechada, apenas com uma abertura. Quando colocamos nessa abertura um papel com uma questão, recebemos, pouco depois, o mesmo papel com a resposta escrita. Se a questão não fizer sentido a resposta será uma coisa do estilo "Isso não faz sentido!" ou "Importa-se de reformular a pergunta?". Também podemos pôr ofensas, piadas ou questões a que ninguém sabe responder como "Será o Porto Penta-Campeão Nacional no próximo ano?" Para todas estas questões receberíamos respostas sensatas e inteligentes. Suponhamos ainda que aquilo que estava dentro dessa caixa poderia memorizar o nosso nome, e as coisas que já tínhamos relatado e as associava a novas situações lembrando-nos de antigas situações e comparando-as com as novas conforme fosse apropriado. Provavelmente julgaríamos que, dentro da caixa estaria alguém ou algo inteligente, talvez um sábio ou um ouvinte atento e bondoso que nos responde de um qualquer sítio da Internet.[16]

Dentro de pouco tempo (alguns séculos) é possível que existem realmente caixas como esta, onde as perguntas sejam feitas oralmente ou por outro processo qualquer, embora o que esteja dentro da caixa seja provavelmente apenas um conjunto imóvel e desinteressante de *chips* de silício que nos responde como se nos conhecesse melhor que nós próprios. É uma hipótese que uma mente do século XX não pode contemplar, certamente, sem alguma perturbação. Pensar que este facto, que ainda não aconteceu, é possível, leva ao choque de duas convicções profundas: *i*) as máquinas não têm mente e *ii*) a inteligência exige uma mente. Uma resposta congruente seria dizer que as máquinas não *são* inteligentes. Mas esta resposta deixará de ser intuitivamente convincente à medida que os computadores deixam o estado arcaico que agora têm (semelhante ao dos primeiros automóveis). Imaginemos que os computadores exibirão de facto um comportamento inteligente, em muitos aspectos indestinguível do comportamento inteligente de uma pessoa, e noutros casos, exibindo até o que, numa pessoa, seria considerado uma maior rapidez ou profundidade de raciocínio. Neste caso, o que pode significar a frase 'os computadores não são inteligentes'?

[15] Esta explicação da experiência do Quarto Chinês parte do princípio que o leitor está já bem familiarizado com ela.

[16] Para uma conversa que parece revelar alguma inteligência, embora sobre uma temática muito reduzida, ver um excerto de uma conversa com SHRDLU, Hofstadter, D., *Gödel, Escher, Bach…*, pp. 586-593.

Note-se que aquilo que gostaríamos de poder dizer é que os computadores, apesar de exibirem todos os sinais de um comportamento inteligente, não são inteligentes (não pensam). Uma maneira simples de fazer isto é dizer que os computadores, sendo meras máquinas, são cegos ao significado dos próprios projectos que executam. Eles exibem a inteligência do programador que se repete num meio estável e previsível. São como textos num livro que se reescrevem a partir das palavras já impressas. Um livro não sabe o que tem lá escrito. Tem uma forma mas não têm uma semântica, não tem intencionalidade. Da mesma forma todos os comportamentos de um computador poderiam ser descritos como exibindo o comportamento inteligente inculcado pelo programador.

Mas, como Turing e outros notaram,[17] nós não temos maneira de saber se algo ou alguém tem, ou não, estados mentais. Eles são 'subjectivos',[18] invisíveis para todos menos para aquele que os possui. Além disso, os cérebros dos animais parecem ter um funcionamento não aleatório, semelhante ao de uma máquina e mesmo assim nós temos intencionalidade. Portanto, negar que os computadores tenham estados mentais só porque são máquinas, parece-se demasiado com afirmações de alguns filósofos de há mais de um século (como Descartes), segundo as quais os animais não tinham sensações ou consciência apesar de se comportarem como tal. Repare-se bem, eles pareciam ter sensações, comportavam-se como se as tivessem, mas na verdade não as tinham, eram simples máquinas. Tentou-se também, com um grau variável de sucesso, aplicar o mesmo argumento a uma ou outra raça humana.

Por isso não é de estranhar que Turing responda assim ao argumento de que os computadores não podem *sentir* certas coisas:

> "O que é importante nesta incapacidade [de apreciar morangos com natas] é que ela contribui para algumas outras incapacidades, *e.g.*, para a dificuldade de ocorrer o mesmo tipo de amizade entre homem e máquina que [ocorre] entre homem branco e homem branco e homem negro e homem negro."[19]

[17] Turing, "Can a Machine Think?" in *The World of Mathematics*, Tempus, 1988, IV volume, p.2087. Este artigo foi publicado pela primeira vez em 1950 com o nome de "Computing Machinery and Intelligence" na revista *Mind*, LIX, pp. 433-460 e está parcialmente disponível na Internet.
[18] No sentido em que Searle utiliza o termo V., *Mente, Cérebro e Ciência*, p.21.
[19] Turing, op.cit, p.2087.

A decisão de atribuir a consciência a uma máquina não poderia ser dada tendo a certeza de que a máquina tinha consciência. Mas do mesmo modo a decisão de atribuir consciência ou intencionalidade a um homem, a uma mulher, a um negro ou a um golfinho não poderia ser dada de acordo com a certeza. Como Turing afirma:

> "[A] única maneira através da qual alguém pode estar **seguro** de que uma máquina pensa [conscientemente] é *ser* a própria máquina e sentir-se a si próprio a pensar. Poderia então descrever esses sentimentos ao mundo, mas é claro que ninguém ligaria. Da mesma forma, de acordo com esta perspectiva, a única maneira de saber que um homem pensa é ser esse homem particular."[20]

Portanto, a conclusão é que a decisão sobre se os computadores (e os animais e os homens) têm mentes teria de ser dada com base, não na certeza, mas noutra coisa qualquer. Como não há qualquer critério seguro a resposta de Turing é: o comportamento das coisas. Se elas se comportarem inteligentemente teremos de supor que são uma espécie de bichos inteligentes. Se não fizermos essa suposição, estaremos a fazer uma distinção arbitrária (tão arbitrária como as que se faziam entre homens brancos e negros.) baseada em hipóteses não demonstradas: é impossível *ser* o computador (tal como é impossível ser outro homem).

Era neste pé que as coisas estavam quando Searle propôs a sua experiência do Quarto Chinês. O grande relevo que tem sido dado à experiência de Searle é, pensamos, devido ao facto de se debruçar sobre a questão principal que levou ao desenvolvimento do behaviourismo: há uma maneira segura de verificar a existência da intencionalidade?

Imaginemos novamente a caixa fechada, mas desta vez Searle está lá dentro. A caixa continua a receber perguntas e a dar respostas. E o que Searle faz é responder às questões utilizando, não as suas capacidades humanas, mas os métodos que os computadores têm ao seu dispor. Para evitar que Searle utilize as suas capacidades mentais vamos escrever as perguntas numa língua que ele não conhece, neste exemplo o chinês. Searle disporá então de um conjunto de livros (o programa) que lhe dirão o que ele deve fazer. Por exemplo, quando aparecer um determinado símbolo chinês ele deve olhar para o seu aspecto, compará-lo com os símbolos desenhados nos livros e

[20] Turing, op.cit. p.2086. Negrito da nossa responsabilidade.

obedecer às instruções que os livros contém para aquele símbolo. Normalmente os livros dirão a Searle para colocar um outro símbolo à saída do quarto. Suponhamos que os símbolos que aparecem no exterior da caixa são as respostas apropriadas às questões colocadas em chinês. Vemos portanto que o computador – Searle – exibiu um comportamento inteligente sem compreender o significado de nenhum dos símbolos chineses. Tudo o que o computador – Searle – sabe é cumprir à letra as instruções do programa. Portanto, da execução de um programa não se pode apreender os significados que podem ser apreendidos do programa por um observador exterior.

O objectivo da experiência de Searle consiste, em síntese, em atribuir intencionalidade a um computador e ver se, a partir da execução do programa, ele poderia compreender os conteúdos que os seus símbolos representam para um observador exterior. Enquanto que Turing tenta derivar o 'interior' da máquina a partir do seu comportamento exterior, Searle imagina-se ele próprio dentro do computador. A ideia é que, como Searle-dentro-da-caixa é uma máquina de Turing, e não compreende chinês apesar de se comportar como se compreendesse, então o CPU que também é um máquina de Turing, também não precisa de compreender chinês só porque se comporta como se o compreendesse:

> "se a efectivação do programa apropriado do computador para a compreensão do chinês não é suficiente para nos *dar* uma compreensão do chinês, então também não basta para dar a qualquer outro computador digital *uma compreensão* do chinês."[21]

Por outras palavras, se puséssemos um qualquer computador digital que soubesse inglês no lugar de Searle (por exemplo outra pessoa ou um computador de silício vindo do futuro) ele não precisaria de saber chinês para funcionar correctamente. É indiferente se estamos na presença de um computadores biológicos ou feitos noutro material, intencionais ou não, conscientes ou não.[22] Seria impossível, fosse qual fosse

[21] Searle, op.cit. p.41.

[22] Também é indiferente a complexidade e eficácia do programa podemos admitir que um computador consiga, no futuro, simular as capacidades humanas até aos mais ínfimos pormenores e em todas as áreas. A única coisa que o argumento rejeita é que esse computador tenha 'conteúdos mentais', é apenas isso que está em causa. Assim, mesmo que um computador fosse capaz de discutir argumentos filosóficos ou resolver problemas da física atómica, de se rir com algumas anedotas e se mostrar consternado nas ocasiões certas, poderíamos dizer que 'a consciência, os pensamentos, os sentimentos, as emoções e tudo o resto' (op.cit.p.46) estariam ausentes.

o sistema computacional, perceber os conteúdos dos símbolos chineses se apenas tivéssemos os símbolos. E as regras de manipulação dos símbolos, como se vê no exemplo, não são suficientes para nos dar o seu conteúdo. Portanto, como os computadores de silício só operam símbolos segundo regras, não podem atribuir-lhes qualquer conteúdo.[23] *QED.*

[23] Ou seja, se um dia conseguirmos pôr um computador a *compreender* o inglês para substituir Searle dentro da caixa, ele teria – segundo este argumento – de alcançar a semântica do inglês de um modo não computacional. (Cf. Searle, *Mentes, Cérebro e Ciência,* pp. 49-50: "as propriedades computacionais do cérebro não são simplesmente suficientes para explicar o seu funcionamento para produzir estados mentais.") Os computadores simplesmente sintácticos não compreendem sequer o significado da sua própria linguagem. Limitam-se a manipular símbolos não interpretados.

3. Críticas ao argumento de Searle.

As críticas principais feitas a Searle são quatro: a crítica do sistema, do robot, do simulador[24] e a combinada. Destas críticas a que nos parece mais central é a crítica do sistema. De facto é dela que dependem as outras críticas que partem do princípio de que é a execução do programa do seu conjunto que permite criar a intencionalidade. Imaginemos o seguinte contra-exemplo ao argumento de Searle:

> "Imaginemos que as sinapses no interior de um cérebro de um ser humano, por qualquer razão,[25] deixavam de responder aos disparos dos axónios. Para conseguir manter vivo esse ser púnhamos em cada uma das sinapses um minúsculo ser humano que tinha como função fazer disparar a sinapse de acordo com o nível de excitação e o valor que define o disparo. O nível de excitação poderia ser considerado o *input* e o valor que define o disparo seria definido através de uns livros escritos em inglês e distribuídos a cada um dos liliputianos que fariam funcionar esse grande cérebro. Ora, parece claro que nenhum dos liliputianos saberia o que Searle estaria a pensar num determinado momento, ou a fazer; tudo o que eles sabiam era cumprir as instruções dos livros ingleses que faziam corresponder, a cada nível de excitação, um determinado comportamento: disparar ou não disparar a sinapse. Portanto, se nenhum dos liliputianos (nem todos no seu conjunto) compreende o significado de cada sinapse também as sinapses se limitam a responder a estímulos, a manipular símbolos (se descritas enquanto parte de um computador digital), sem compreender o seu significado."[26]

Porque é que este exemplo não é revelador? Talvez porque ninguém defende que as sinapses isoladas deveriam ser conscientes ou ter intencionalidade. Mesmo Searle afirma que

> "embora possamos dizer de um cérebro particular: «Este cérebro é consciente», ou «Este cérebro sente sede ou dor», nada podemos dizer de algum neurónio particular no cérebro: «Este neurónio tem dor, este neurónio sente sede.» … Nada é mais comum na Natureza do que serem as características de superfície de um fenómeno causadas por e

[24] Introduzida mais recentemente pelos Churchland em "Could a Machine Think?", in *Scientific American,* 1990, 262.
[25] Poderia ser uma redução de temperatura ou uma desordem química.
[26] Este exemplo é um «original» deste trabalho.

realizadas numa microestrutura, e essas são exactamente as relações exibidas pela conexão da mente ao cérebro."[27]

Poderíamos portanto afirmar que, no exemplo dos liliputianos, o que se exibe é a ausência das propriedades da macro estrutura na micro estrutura, o que é algo trivial e sem consequências. Muitos defensores da IA forte julgam que é aí que reside a fraqueza do argumento de Searle. De facto, afirmam esses críticos[28] que ninguém supõe que o CPU adquira consciência por correr um programa. Nem que seja o programa isolado que possua intencionalidade. Pelo contrário, no melhor dos casos, é o funcionamento integrado de todo o sistema que pode atingir a intencionalidade. Esta crítica é de algum modo corroborada pela dificuldade de adequar o exemplo de Searle a esta crítica sem lhe retirar o carácter intuitivo. Se Searle, em vez de ser o CPU fosse todo o sistema teria, não de estar dentro da caixa recebendo símbolos, lendo livros e efectuando as operações neles descritas, ele seria em vez disso a própria caixa com tudo o que está lá dentro. Se o programa fosse realmente muito bom teríamos de imaginar Searle – sem caixa, nem programa, nem CPU distintos – a responder a perguntas chinesas num chinês perfeitamente correcto e imaginar ao mesmo tempo que Searle não compreenda uma única palavra de chinês.

Embora em *Mente, Cérebro e Ciência* Searle não responda cabalmente à crítica do sistema,[29] noutras passagens Searle reformula a sua experiência de uma maneira que a torna intuitivamente atraente mas que torna os adeptos da IA descontentes. A ideia é

"Primeiro, em vez de o termos a memorizar uma biblioteca, devemos imagina-lo a *memorizar* toda a biblioteca. Segundo, em vez de escrever notas em papeis [os símbolos chineses de saída], teria de memorizar o que os papeis diriam [substituindo os papeis pela voz e pela audição]. … Mas como Searle enfatizaria, embora pareça aos falantes de Chinês que estaria a conduzir um discurso inteligente [*learned*] com eles em chinês,

[27] Searle, op.cit., p.28.
[28] Por exemplo, Block, Hofstader ou Stower.
[29] Cf. p.42. A única razão que Searle apresenta é uma afirmação: tal como o CPU não consegue passar da sintaxe para a semântica "também não o consegue todo o sistema". Note-se que Searle limita-se a repetir A3, mas a base de A3 encontra-se no próprio exemplo que é a base da refutação da IA forte (Cf. primeira citação de Searle no nosso trabalho). Portanto, se o exemplo perde a força quando considerado o conjunto do sistema é todo o argumento, e não apenas os limites da experiência conceptual, que é abalado.

> tudo o que estaria consciente [*aware*] de fazer é pensar acerca dos sons que o programa diz para fazer a seguir."[30]

O problema deste exemplo é que não é claro se a memorização dos livros faz com que Searle seja os livros. E se Searle não for o sistema então não responde às objecções desta crítica. Exemplificando, podemos reformular o exemplo dos liliputianos imaginando alguém que memorize todos os livros dos liliputianos e realize mentalmente as operações dos liliputianos. Como é óbvio, mesmo pensando que o cérebro que os liliputianos operam é consciente, quem imagina todos os biliões de liliputianos não vê senão cada uma das operações em particular. Não se passa, como Searle afirma, do nível sintáctico para o semântico, nem é possível passar deste modo, porque cada operação continua a valer como se estivesse isolada.

Poderíamos colocar também este exemplo em relação à humidade da molécula de água. Enquanto imaginássemos o comportamento de cada molécula individual de água não conseguiríamos sequer perceber o que é a humidade. Esta só surge quando se enquadra a molécula no conjunto. Quando cada molécula desaparece para dar lugar ao conjunto.

Quando Searle se imagina a memorizar todos os livros ele mantém a descrição, não ao nível do conjunto, de um único sistema composto de muitas partes indestinguíveis, mas de muitas partes que tornam o conjunto invisível. Portanto, para um defensor da perspectiva forte da IA iria parecer que, no exemplo de Searle, embora Searle 'implementasse' um «sistema intencional» não poderia aperceber-se da intencionalidade do sistema. Tal como uma descrição atómica completa do cérebro não nos daria os conteúdos mentais possuídos por esse cérebro.[31]

[30] Ned Block, "The Mind as the Software of the Brain", cap. 4. Para outra descrição desta experiência de Searle ver Martin Stower, *Searle'Searle Chinese Room Argument*, §6.1.

[31] Para isso precisamos de criar correlações que *não* são dadas pela (nem podem ser derivadas da) simples descrição dos movimentos dos átomos no interior do cérebro, seja ela tão rigorosa ou abrangente quanto o quisermos. Essas correlações seriam, por exemplo, afirmar que à excitação de certas áreas corresponde um determinado conteúdo mental.

4. Conclusão

Searle responde aos adeptos da IA forte mostrando que uma descrição de baixo nível de um computador não revela qualquer intencionalidade e com o 'axioma' de que, se a semântica não existe ao nível mais baixo, então também não pode surgir nos mais elevados. No entanto, o mesmo tipo de objecção pode ser levantado à teoria de que é o cérebro que causa as mentes. Numa descrição de baixo nível (à escala dos átomos ou dos neurónios, por exemplo) não aparece nem a intencionalidade nem a consciência. E no entanto a nossa experiência quotidiana e as nossas teorias científicas parecem afirmar, conjugadas, que é de facto possível fazer essa passagem. Se aceitamos que as propriedades de um sistema não se resumem às propriedades conjugadas das suas partes, não poderíamos também imaginar que A3 pode, em quantidades inimagináveis de complexidade, ser falsa?

Este argumento tem a particularidade de ser paralelo ao argumento que Searle tinha usado para defender que o cérebro, sendo puramente físico num certo nível de descrição microscópica, poderia possuir, ao nível macroscópico, de articulação de muitas moléculas, características bastantes diferentes das apresentadas ao nível microscópico. Assim 'não podemos pegar numa molécula de água e dizer esta aqui está húmida'. Este facto, que é trivial e costuma ser referido pela máxima 'o todo é maior do que a soma das partes', poderia ser a analogia exacta do que os funcionalistas pretendem fazer ao nível computacional. Funções simples e não intencionais podem, integradas num sistema mais complexo, que integra grandes quantidades de informação, estruturadas de forma adequada, permitem criar um 'conhecimento' acerca do bastante amplo acerca do mundo

No argumento original de Searle podemos distinguir uma tese negativa «os computadores não pensam» e uma positiva: «os cérebros são a causa do pensamento». No nosso trabalho analisámos apenas a experiência conceptual do Quarto Chinês, que corresponde apenas à parte negativa do argumento de Searle. Tomada como uma refutação da posição behaviourista de Turing ela é suficientemente eficaz. No entanto, é claro pela caracterização que Searle faz da 'IA forte', que o argumento visa atacar o funcionalismo. Mas, como o argumento de Searle contra o funcionalismo pode ser visto como a defesa do reducionismo ao nível dos computadores digitais, e, como

Searle defende uma tese oposta no que toca à relação entre cérebros e computadores, seria preciso encontrar uma diferença entre os computadores de silício e os neurónios de água e carbono que justificasse a diferença de tratamento.

Ora é neste ponto que o argumento de Searle parece mais frágil. Porque Searle afirma que "as propriedades computacionais do cérebro não são suficientes para explicar o seu funcionamento para produzir estados mentais"[32] mas a característica que Searle acrescenta é que "os cérebros são máquinas biológicas."[33] Ora a biologia só por si não é suficiente para explicar nada. Se isto fosse realmente importante seria, dentro de alguns anos, possível criar computadores a partir de conjuntos de neurónios pré-programados. Mas é evidente que nem toda a actividade neural é consciente. Uma grande parte dos neurónios do cérebro estão ocupados com operações como o equilíbrio. E tal como Roger Penrose faz notar a consciência é uma capacidade do cérebro que só é útil e utilizada em circunstâncias relativamente raras como a aprendizagem.[34] Por outro lado, a maior parte das coisas que fazemos (como escolher as teclas num teclado ou as palavras num discurso) não utilizam a consciência nem incluem a intencionalidade.[ii] O que parece ser um argumento forte contra o funcionalismo já que parece que não basta a activação de um programa para originar a intencionalidade.

Portanto, em relação às pretensões por nós esboçadas no início deste trabalho diríamos que o argumento de Searle é evidente se aceitarmos uma certa forma de reducionismo. No entanto, como, com base no reducionismo, somos incapazes de explicar a existência dos nossos próprios estados mentais aceitando ao mesmo tempo o realismo ingénuo,[35] e como Searle aceita o realismo ingénuo, temos de abandonar o reducionismo. Sem uma base sólida para distinguir cérebros de computadores não biológicos esta ambivalência em relação ao reducionismo aparece como a pedra de toque (injustificada) de toda a argumentação de Searle.

---

[32] Op.cit. p.50.
[33] Idem.
[34] Cf. *The Emperor's New Mind*, pp.405-431, esp. pp.409-413.
[35] Como foi notado por vários críticos, Searle de facto não consegue explicar como é que se passa do nível físico para o nível mental. Talvez daí o recurso à expressão 'a melhor maneira de mostrar que algo é possível é mostrar que efectivamente existe', o que é perfeitamente aplicável aos programas de IA. (Por exemplo à diferenciação entre informação e conhecimento no interior de um programa.)

Por outro lado parece haver uma incongruência entre o que Searle afirma que fez 'refutei a ideia de que o cérebro é um computador digital' com a afirmação várias vezes proferida de que o cérebro pode ser descrito como um computador digital. Parece que o que Searle está a dizer é que, embora aceite que o cérebro pode ser descrito como um computador digital ele não é apenas um computador digital. E, mais importante, são estas outras capacidades que permitem ao cérebro criar a intencionalidade. No entanto, esta conclusão é também a premissa de todo o argumento. Porque, se partíssemos da proposição inversa estaríamos a dizer que o reducionismo não é válido ao nível dos computadores digitais complexos e portanto, concluiríamos do argumento de Searle que A1 ou A3 (ou ambos) estão errados. O argumento de Searle seria muito mais evidente se não houvesse consciência humana.

A questão, portanto, que parece poder decidir o argumento de Searle é precisamente a de saber se o *funcionamento* do cérebro pode ser descrito e previsto como um computador digital (isto é se pode ser previsto por uma máquina de Turing). Esta questão talvez se torne uma questão empírica dentro de alguns séculos mas por enquanto só podemos especular sobre a possibilidade de algum computador digital sem os «poderes não computacionais» do cérebro poder duplicar o pensamento humano.

No entanto, nesta perspectiva consideramos que A3 não é uma evidência lógica. Se considerarmos que A3 é um axioma da mesma forma que as leis de Leibniz são axiomas podemos de facto considerar, como Searle faz, que o seu argumento é uma verdade lógica e que os cérebros devem portanto ter outros poderes que não os computacionais. A resposta dos críticos de Searle (através da crítica do sistema) parece ser uma maneira clara de mostrar que A3 não pode ser considerado como uma verdade lógica. [iii]

5. Apêndice 1.

A ideia de que

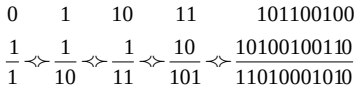
"Os zeros e uns … são simples numerais; nem sequer estão em vez de números"[36]

não é inteiramente evidente por várias razões. Uma delas é que existe, para cada contexto, um número indefinido de sequências de símbolos realmente sem sentido, ou seja, são incapazes de representar seja o que for. Por exemplo, na língua portuguesa, a sequência de símbolos 'xjiwjoa' não tem qualquer significado. Em geral há, para cada contexto, uma quantidade gigantesca de sequências de símbolos sem sentido quando comparada com o número de sequências de símbolos disponíveis com sentido. Portanto, pareceria plausível caracterizar 'xjiwjoa' como uma sequência sem significado e 'carro' como uma sequência com significado. Imaginemos um livro com a palavra carro que só é lido uma única vez. O facto de alguém ler nesse livro a palavra 'carro' num momento t1 não faria com que a palavra só tivesse significado em t1. Mesmo depois de lida, a palavra escrita deveria continuar a ter significado.

Poderíamos pensar que, por exemplo, uma máquina de calcular utiliza símbolos que são, de facto, números. E seria por isso que qualquer máquina de calcular pode fazer um número quase infinito de operações matemáticas correctas com um pequeno número de regras.

Por exemplo suponhamos que queremos adicionar os números 2 e 3. Em linguagem binária 2 escreve-se '10' e 3 escreve-se '11'. A adição de dois números binários faz-se de uma forma extremamente simples. Tal como numa adição normal comparamos os dois algarismos com a mesma casa e, no caso de dois 0's mantém-se o valor, no caso de termos um 0 e 1 o valor passa a 1, no caso de termos dois 1's o valor passa a 0 e acrescenta-se 1 à soma da casa seguinte ('e vai um'). Assim 10 + 11 seria igual a 101, que em linguagem decimal quer dizer '5'. Outros exemplos de operações de soma em linguagem binária seriam:

[36] Idem.



Ora, parece evidente que estas sequências de zeros e uns não sabem que representam números, nem o computador sabe que elas apontam para o nome ‘5’ ou para o número que esse nome representa. Parece certo que estas sequências não *representam* números. Mas isso não quer dizer que elas não *sejam* números na medida em que podem ser correctamente operadas pelo programa. De certa forma, poderíamos supor que a sintaxe pode, em condições apropriadas, constituir os objectos abstractos que os nossos nomes designam.[37]

[37] Nós também não fazemos contas com os nomes que designam os números, estejam eles escritos em russo, árabe, romano ou grego. Só sabemos que VI mais XXXII é igual a IIXL porque sabemos operar os próprios números.

6. Apêndice 2

Como é que sabemos que os programas são puramente sintácticos? Em *Mente, Cérebro e Ciência,* Searle considera que isto é uma questão da própria definição do que é um programa de computador. Mas ela não é a única definição que podemos dar, e é um facto que podemos pensar que certos símbolos, no contexto do programa, têm uma semântica. Portanto, A1 teria uma base muito frágil. Seja como for não devemos pensar que o argumento do Quarto Chinês pode justificar A1 (ele destina-se a reforçar apenas A3).

Isto vê-se facilmente se pensarmos que a experiência de Searle mantêm a sua persuasão intuitiva quer A1 seja verdadeiro ou falso. Mesmo os defensores da IA forte admitem que há muitos programas que não têm uma semântica capaz de compreender o chinês (por exemplo um processador de texto ou uma versão chinesa do ELIZA). Portanto, o exemplo de Searle seria possível se imaginássemos um programa adequado (respostas correctas mas raciocínio insuficiente).

Num programa do tipo ELIZA[iv] certamente que haveria simulação sem duplicação. Neste caso os livros dentro do quarto chinês abordariam apenas os aspectos sintácticos dos símbolos chineses. Mas programas mais complexos, que envolvessem análises da linguagem não só sintácticas mas semânticas, poderiam certamente atingir a compreensão do chinês.[38] Assim alguém que recusasse A1 poderia dizer que aceitava intuitivamente o exemplo de Searle com as suas implicações. Mas, neste caso, poderíamos dizer que o exemplo de Searle só se aplica em casos em que os livros que Searle-dentro-do-quarto utiliza para dar as respostas não lhe explicam o que é o chinês. Poder-se-ia então encontrar facilmente um contra exemplo. Bastaria imaginar um quarto chinês que, em vez de livros com regras estivesse cheio de livros que ensinassem a ler chinês, que atribuíssem a cada símbolo chinês uma certa entidade e nos desse uma imagem dos chineses e da sua cultura, mentalidade, história, etc.

[38] É à construção deste tipo de programas que a IA, como disciplina, aspira. Na verdade é um processo análogo a este que os programas de tradução mais complexos utilizam. Cf. *Artificial Intelligence*, esp. cap. 15).

Neste caso o que está em causa (respeitando a validade do Argumento do Quarto Chinês mas não A1) são as limitações desse programa e não do computador enquanto tal. Mesmo interpretado desta maneira o exemplo de Searle pode ser útil porque põe em causa os testes puramente behaviouristas de inteligência, dissociando o *output* correcto do raciocínio correcto.

Mas é claro que o objectivo de Searle não é por em causa o carácter behaviourista do teste de Turing. Pelo contrário, o ataque de Searle é dirigido contra o funcionalismo, e dirige-se portanto, principalmente, contra os programas de IA mais complexos e eficazes que podemos esperar vir a alcançar.[39] Já que, segundo o funcionalismo, são estes que, estando mais próximos das funções realizadas pela mente humana, teriam mentes mais parecidas com as humanas.

Poderíamos pensar que Searle mostra a validade de A1 porque demonstra que, independentemente da complexidade e abrangência do programa, o CPU (a parte que Searle-dentro-do quarto desempenha) nunca pode compreender senão os aspectos sintácticos dos *inputs*. Tudo o que ele compreende são símbolos 'esticados' e 'encolhidos':[40] os uns e os zeros. Ora no exemplo, Searle-dentro-do-quarto compreende as instruções que lhe são dadas em inglês e o aspecto dos símbolos chineses, o que viola A1. Ora, se Searle-dentro-da-caixa compreende o inglês, se (respeitando a analogia) o CPU compreende linguagem máquina, então o que nos impediria de imaginar um CPU que compreendesse a semântica do chinês?[v]

[39] Note-se que o exemplo que utiliza no seu artigo de 1980 é o de um programa que utiliza *scripts*, que são formas complexas de introduzir semântica num programa (sobre o modo de introduzir semântica num programa Cf. *Artificial Intelligence*, esp. cap. 15). Além disso, mesmo em *Mente, Cérebro e Ciência*, Searle admite que pode em princípio haver programas capazes de simular o comportamento de um ser humano sem quaisquer restrições.
[40] Searle, op.cit., p.40.

[i] Larry Hauser, numa tese de doutoramento divulgada na Internet e também num artigo, defende que A1 é falsa quando o programa é corrido: "Para ver a falsidade de A1, basta ver que só programas a *correr* ou a execuções de programas são candidatos pensantes [*candidate thinkings*], e.g., de acordo com a formulação de Newell da hipótese funcionalista (que o próprio Searle cita), "a essência do mental é a *operação* de um sistema simbólico físico." [itálicos de Hauser] ninguém supõe que instanciações inertes (não executadas) de Programas (e.g. em disquetes), "por elas próprias", pensem ou sejam suficientes para o pensamento." in "Searle's Chinese Box: Debunking the Chinese Room Argument", http://members.aol.com/lhauser2/chinabox.html, também publicado em *Minds and Brains, 7*: 199-226, 1997, Capítulo 5 (a meio).

A ideia de Hauser parece ser a de que o impulso eléctrico no computador adquire, de um modo tão misterioso como os impulsos eléctricos no cérebro, uma certa semântica. Hauser parece também pensar que mesmo sistemas simbólicos relativamente simples possuem já uma semântica ("até as vulgares calculadoras de bolso têm realmente propriedades mentais". V. cap. 6). Note-se ainda que Hauser aceita a validade de A3.

[ii] Podemos conduzir durante horas (quando pensamos noutra coisa) sem termos presente a mínima compreensão do que é conduzir, do sítio em que nos encontramos, dos carros que nos rodeiam. É claro que as imagens continuam a ser enviadas ao cérebro e algo aí determina a colocação das mudanças, trava, vira o volante, liga as luzes, para nos semáforos, lembra-se do caminho, etc. Mas a nenhuma dessas imagens é atribuído um significado, um conteúdo mental, elas parecem não ter semântica. Só no caso de uma situação de emergência a consciência é chamada a intervir. Isto pode ser considerado como uma refutação do behaviourismo se considerarmos que conduzir um carro desta forma é um acção inteligente, mas parece aplicar-se mais ao funcionalismo. Porque, neste caso parece que aplicamos o programa correcto para conduzir mesmo sem termos uma compreensão do que estamos a fazer. Provavelmente um funcionalista diria que estaríamos a implementar um programa de que não tínhamos consciência, mas que as imagens dos carros têm, para ele (o programa), realmente uma semântica. Por outro lado, parece também pôr em causa a afirmação de Searle: "Os nossos estados mentais internos têm, por definição, certos tipos de conteúdos." Op.cit. p.39.

[iii] Como curiosidade e para efeitos de referência futura notamos que o exemplo da caixa chinesa apareceu primeiro em 1980 com o artigo "Minds, Brains, and Programs", in *Behavioral and Brain Sciences* 3, pp. 417-424. Este artigo é publicado juntamente com várias objecções às quais Searle responde nas pp. 450-456, no artigo, "Intrinsic Intentionality".

Dois anos mais tarde Searle responde a algumas críticas no artigo "The Chinese Room Revisited", *Behavioral and Brain Sciences* 5, pp. 345-348.

O último capítulo de *Intentionality: an Essay in the Philosophy of Mind*, New York: Cambridge University Press contém também uma versão reformulada da experiência da caixa chinesa, que reaparece em 1984, no segundo capítulo de *Minds, Brains, and Science*, Cambridge: Harvard University Press.

Outras formulações da experiência são dadas em

Searle, J. R. (1988), "Minds and Brains Without Programs", in C. Blakemore and S. Greenfield, eds., *Mindwaves*, Oxford: Basil Blackwell, pp. 209-233.

Searle, J. R. (1989a), "Reply to Jacquette", *Philosophy and Phenomenological Research* XLIX, pp. 701-708.

Searle, J. R. (1990a), "Is the Brain's Mind a Computer Program?", *Scientific American* 262, pp. 26-31.

Outros textos de Searle relacionados:

Searle, J. R. (1987), "Indeterminacy, Empiricism, and the First Person", *Journal of Philosophy* LXXXIV, pp. 123-146.

Searle, J. R. (1989b), "Consciousness, Unconsciousness, and Intentionality", *Philosophical Topics* XVII, pp. 193-209.

Searle, J. R. (1990b), "The Causal Powers of the Brain", *Behavioral and Brain Sciences* 13, p. 164.

Searle, J. R. (1990c), "Who is Computing with the Brain?", *Behavioral and Brain Sciences* 13, pp. 632-640.

Searle, J. R. (1991). "Perception and the Satisfactions of Intentionality", in E. Lepore and R. Van Gulick, eds., *John Searle and His Critics*, Cambridge, MA: Basil Blackwell, pp. 181-192.

Searle, J. R. (1992), *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Searle, J. R., J. McCarthy, H. Dreyfus, M. Minsky, and S. Papert (1984), "Has Artificial Intelligence Research Illuminated Human Thinking?", *Annals of the New York City Academy of Arts and Sciences* 426, pp. 138-160.

[iv] ELIZA é um programa que simula um programa de inteligência artificial. Em comparação com um programa de IA, é fantasticamente simples. Limita-se, na maior parte das vezes, a pegar nas frases do interlocutor e a repeti-las com ligeiras alterações. Por vezes também introduz frases pré-programadas mas no geral o programa não tem sequer um dicionário, é praticamente desprovido de noções sintácticas da linguagem (também chamado *parsing*) e não tem qualquer semântica. Um input do estilo "oiew dof para iwe fiup" teria uma resposta do estilo "fala-me mais de "dof para iwe fiup". No entanto, numa conversa normal é normalmente muito difícil perceber que é um computador que está do outro lado (no *chat*, por exemplo, estes programas são praticamente indestinguíveis dos interlocutores humanos).

[v] As restrições neste caso teriam apenas a ver com o número de palavras diferentes que o CPU fosse capaz de reconhecer, ora os CPU's têm uma linguagem que é veiculada em símbolos binários mas não é binária. Um processador vulgar actualmente processa símbolos, cada um dos quais constituído por uma sequência de 64 *bits* (zeros ou uns), o que corresponde a um vocabulário máximo de 2^{64} palavras. Ou seja o CPU mais barato que se pode encontrar numa loja de computadores pode memorizar mais palavras do que aquelas que contêm todos os dicionários europeus no seu conjunto (Note-se que os CPU's actuais reconhecem apenas um pequeno número de instruções, mas a linguagem em que falam não é, certamente, binária.).