Alguns Problemas de Filosofia da IA
Sofia Miguens

O objecto de investigação da IA é a natureza da cognição, e ter uma teoria da cognição – nomeadamente de fenómenos como a categorização e identificação de objectos, a resolução de problemas, a decisão, a consciência — sempre foi a ambição dos filósofos. Por isso, desde os inícios da disciplina, os filósofos interessam-se pelas investigações em IA e pelo que elas podem mostrar quanto à inteligência humana e quanto à inteligência em geral.

Curiosamente, os filósofos têm adoptado posições extremas quanto à IA, desde a defesa de uma impossibilidade por princípio da criação de inteligência e consciência não naturais, por razões várias, até á convicção de que através da IA poderá surgir uma concepção mais geral e mais abstracta sobre a natureza da inteligência, que coloque os humanos e todos os seres biológicos inteligentes como apenas casos particulares de um fenómeno geral. Basicamente, a ideia é que se qualquer sistema com o tipo correcto de organização funcional pode ser inteligente e até mesmo consciente, e se essa condição pode ser formulada independentemente da matéria de que fôr constituído o sistema e independentemente das suas origens, outros sistemas que não os humanos e outros seres biológicos poderão, pelas mesmas razões que estes, ser inteligentes e conscientes.

A crítica filosófica à IA: H. Dreyfus e J. Searle Hubert Dreyfus (http://socrates.berkeley.edu/~hdreyfus) e John Searle (http://socrates.berkeley.edu/~jsearle) são dois filósofos com posições críticas face à IA.

Hubert Dreyfus

Hubert Dreyfus foi talvez o primeiro filósofo a notar a sobreposição de interesses da IA e da filosofia. Escreveu como resultado desse encontro textos polémicos nomeadamente um relatório intitulado *Alchemy and Artifical Inteligence* (Rand Corporation, Santa Monica, California, 1965) e depois o livro *What Computers Can't Do* (1972), criticando as pretensões da IA.. Dreyfus parte do princípio de que a IA trabalha os mesmos problemas que a filosofia, problemas como a natureza do entendimento e do conhecimento. A sua crítica insiste sobretudo na relação entre os modelos então desenvolvidos na IA, baseados na ideia de mente como sistema simbólico e na ideia de inteligência como resolução de problemas e a tradição racionalista e intelectualista em filosofia, tradição essa que Dreyfus critica. Para Dreyfus, a IA repetia portanto os erros 'intelectualistas' da filosofia, erros já apontados por filósofos como Heidegger, Wittgenstein e Merleau-Ponty.

A crítica de Dreyfus tinha como centro a ideia de que o pensamento não pode ser considerado como consistindo em representações simbólicas do mundo, nem a mente como um sistema governado por regras ocupado com a resolução de problemas, pois essas definições excluem partes importantes e básicas do mental. As 'partes excluídas' do mental eram para Dreyfus por exemplo os movimentos corporais e o reconhecimento de padrões e estas estariam subjacentes á própria possibilidade das habilidades (*skills*) explícitas envolvidas nas representações e na resolução de problemas.

Os escritos de Dreyfus foram muito mal recebidos, mas apresentavam

uma perspectiva sóbria sobre alguns exageros de previsão que acompanharam os primórdios da IA. Apesar de algumas das previsões do próprio Dreyfus terem sido simplesmente desmentidas (por exemplo a ideia de que um computador não poderia vencer um humano num jogo de xadrez), o que Dreyfus queria criticar era a suposição apriorista não argumentada – defendida, na sua opinião, por exemplo por A. Newell. H. Simon e M. Minsky – segundo a qual regras e representações seriam a natureza da mente. De acordo com o intelectualismo dessa visão até a percepção seria pensamento e resolução de problemas através da manipulação de regras. Esta concepção ocultava o conhecimento do fundo (*background knowledge*) ou de senso comum, que na perspectiva de Dreyfus era fundamental para a própria cognição. O conhecimento do fundo não é conhecimento de factos mas sim, no caso das pessoas, aquilo que elas sabem sem saberem que sabem e que nunca foi aprendido (como por exemplo que as pessoas se movem mais facilmente para a frente do que para trás, num exemplo de Dreyfus, ou que se se entornar água em cima da toalha ela passará para as pernas de quem está por baixo).

Dreyfus insistiu ainda noutras características do pensamento humano, ausentes nas simulações computacionais que então considerava como exemplos: as margens da consciência (*fringe consciousness*), a tolerância da ambiguidade, a ligação a um corpo no mundo que organiza e unifica a experiência de objectos e as impressões subjectivas, a capacidade de aborrecimento, cansaço e perda de motivação e os propósito e interesses que gerem o confronto do sujeito com a situação no mundo, fazendo com que nem todos os factos no mundo sejam igualmente relevantes num dado instante, com que o mundo não seja 'liso'.

O teor das críticas de Dreyfus foi então para cá incorporado no natural desenvolvimento da IA e actualmente Dreyfus admite a proximidade entre os princípios do conexionismo e a tradição anti-racionalista em filosofia.

John Searle John Searle foi conduzido às Ciência Cognitivas a partir do seu trabalho em filosofia da linguagem. É um crítico famoso dos limites do modelo computacional da mente. Searle pensa que as ciências cognitivas são um campo de investigação excitante mas fundado num erro conceptual acerca da natureza da mente, e que a 'hipótese da IA forte' (a ideia segundo a qual qualquer sistema que implemente o programa correcto poderá ter *realmente* uma mente e consciência, logo não será *apenas* uma simulação), é refutável via o Argumento do Quarto Chinês. O Argumento do Quarto Chinês pretende ainda 'refutar' o Teste de Turing (Turing 1950).

O artigo *Minds, Brains and Programs*, onde o argumento é defendido, apareceu na revista Behavioral and Brain Sciences em 1980, juntamente com 28 respostas de críticos. Desde então a experiência mental do Quarto Chinês é incontornável na filosofia da mente, associada à ideia de que não se pode falar de processos mentais sintácticos sem falar de semântica (e, de facto, Searle liga a semânica à consciência).

A experiência mental do Quarto Chinês (QC) consiste no seguinte: alguém, que não fala chinês, está fechado dentro de um quarto onde há símbolos chineses em caixas. Tem um livro de instruções em inglês, que explica como combinar os símbolos chineses e como enviar sequências de símbolos chineses para fora do quarto, quando são introduzidos no quarto outros símbolos chineses, através de uma pequena janela. A pessoa que está dentro do quarto não sabe nada acerca disso, mas as pessoas que estão fora do quarto chamam aos símbolos que introduzem 'perguntas' e aos símbolos que saem 'respostas'. O sistema fala portanto chinês, na perspectiva das pessoas que estão fora. Então, o sistema passa o Teste de Turing, embora a pessoa lá dentro saiba que não percebe uma palavra de chinês. Searle afirma que a experiência mental do QC torna clara a possibilidade de um sistema que tem 'intencionalidade atribuída' mas não

'intencionalidade intrínseca' ou 'semântica genuína'.

O problema é saber exctamente que é que o argumento de Searle mostra. Antes de mais, note-se que o QC é mais propriamente uma parábola, e que posto em forma de argumento daria o seguinte:

Os programas são sintácticos
A sintaxe não é suficiente para a semântica
As mentes têm semântica
Implementar um programa é insuficiente para haver mente

O QC mostraria então que a mente não é um programa e que por isso programar apropriadamente alguma coisa nunca poderia dar-lhe uma mente, já que as propriedades formais não constituem a 'intencionalidade genuína'.

O filósofo David Chalmers parodia este argumento do seguinte modo (Chalmers 1996: 327):

As receitas são sintácticas
A sintaxe não é suficiente para ser-saboroso
Os bolos são saborosos
As receitas não são suficientes para fazer bolos

Chalmers pretende mostrar que o argumento de Searle não distingue entre sintaxe e *implementação* de sintaxe, entre um programa 'na prateleira' e um programa correndo numa máquina física.

O próprio Searle sublinha sempre que o seu argumento não tem nada a ver com um estádio evolutivo particular da tecnologia, mas antes diz respeito a princípios conceptuais. O que o QC pretenderia fazer ver seria que o cognitivismo – no qual é central a ideia de que 'a mente está para o cérebro como o software para o hardware' – está errado ao considerar que não existe nada de *essencialmente* biológico acerca da mente humana (é esta posição que tem como corolário a defesa da IA forte, a ideia, recorde-se, de que qualquer coisa que implementasse o programa correcto poderia ter realmente uma mente. Para Searle, pelo contrário, a mente é essencialmente consciência, e a existência de consciência é um facto biológico.

De facto, na formulação inicial do Quarto Chinês, Searle fala mais propriamente de semântica, e afirma que 'a sintaxe não é suficiente para a semântica'. Mas o que Searle pensa é que não há possibilidade de considerar qualquer coisa como mental a não por relação com a consciência.

A fraqueza da posição de Searle é não dar um critério claro do que se entende por intencionalidade originária ou intrínseca por oposição a intencionalidade atribuída, as noções de que se serve para analisar o QC. O único critério que Searle dá são os poderes causais que alguns sistemas físicos (nomeadamente os cérebros humanos) teriam e outros não.

Variadíssimas objecções foram levantadas ao Quarto Chinês, algumas das quais as a seguir listadas:

a. Com o QC Searle pretende provocar a nossa identificação com o ser humano que simula 'à mão' um programa de IA. Ora, isso não faz sentido pois Searle oculta a questão da escala: um ser humano não poderia simular à mão um programa de IA suficientemente complexo para a tarefa em causa. Isso envolveria meses ou anos, para não falar de horrivel aborrecimento. Searle passa 'por baixo do pano', com o apelo à identificação, uma concepção irrealista da relação entre inteligência e manipulação simbólica, ao impedir que se considere a enorme diferença de complexidade entre o nível do programa e o nível da pessoa e ao sugerir que 'sintamos' a falta de entendimento da

pessoa ao levar a cabo um programa.

b. A resposta básica à situação ao QC é a resposta dos sistemas: é um erro imputar o entendimento ao executor do programa (incidentalmente animado). O entendimento pertence ao sistema como um todo, que inclui os "pedaços de papel" com as regras/símbolos (aliás, nos nossos neurónios também não há entendimento do português que falamos, e no entanto, nós, o sistema global, falamos português, e a falta de intencionalidade genuína dos nossos neurónios não constitui prova da falta de entendimento genuíno em nós)

Vários autores (cf. Hofstadter & Dennett 1981) contrapuseram a Searle, situações semelhantes à do QC, mas em que a nossa intuição vai em sentido contrário ao que se passa no QC:
O psicólogo canadiano Zenon Pylyshyn propôs o seguinte: e se se fosse substituindo célula a célula as células do nosso cérebro por chips programados para manterem exactamente as mesmas funções? Neste caso continuariamos a falar e agir do mesmo modo, mas Searle teria que dizer que deixariamos de significar (*mean*) progressivamente (a nossa interioridade desapareceria de nós, sem ninguém notar, nem nós próprios). O problema é que neste caso não 'tendemos' a ter essa intuição – o que mostra, de novo, Searle não dá critérios para dizer quando é que a intencionalidade genuína está presente ou ausente do sistema.

O filósofos americano John Haugeland propõe a seguinte caso: o cérebro de uma mulher real tem uma lesão, e há um 'agente artificial' que intervem para "conduzir" os neurotransmissores de forma idêntica ao que antes da lesão acontecia naturalmente. O cérebro funciona então exactamente como se a mulher estivesse saudável. Haugeland pergunta a Searle se esta mulher pensaria ou não teria entendimento nenhum do seu próprio pensamento. O próprio Searle concorda que nesse caso devemos olhar para o ponto de vista da mulher e não para o do agente (embora este pudesse, por hipótese ter o seguinte ponto de vista: 'Loucos! Não lhe dêem atenção! Ela é um fantoche cujas acções são causadas por mim!')

O ponto destes contra-exemplos imaginários é mostrar que o QC depende exclusivamente da nossa intuição, e as nossas intuições podem ser feitas variar (é o caso das intuições do próprio Searle). Por isso nenhum bom argumento acerca daquilo a que Searle chama intencionalidade intrínseca deverá depender exclusivamente de intuições.

Ninguém nega que a experiência de Searle é provocadora, sobretudo porque ela diz respeito ao facto de no estudo da cognição estarem sempre em causa vários níveis. Aliás, Hofstadter sugere que o que precisamos para pensar no QC é precisamente do conceito de níveis de implementação, da ideia de que um sistema pode emular outra sistema e, assim sendo, :o 'encaixe' de uma 'máquina' noutra dá origem a 'maquinas virtuais'. Debaixo de cada máquina virtual há sempre outra máquina, sendo apenas num certo sentido a máquina de baixo a 'máquina real'. Em princípio os níveis estão selados uns aos outros (como a pessoa que não fala o chinês que o sistema fala, como os neurónios que não falam o português que nós falamos). Mas os níveis poderiam 'comunicar' e Hofstadter sugere que talvez isto seja o que se passa por exemplo quando um sistema humano aprende uma nova língua, que quando é bem aprendida não corre simplesmente sobre a primeira, como uma espécie de parasita de software, antes deixa de precisar de ser traduzida e torna-se mais fundamental. Hofstadter associa, aliás, esta 'comunicação' entre níveis de um sistema à consciência (que, note, é o verdadeiro problema de Searle no QC).

Em suma, saber o que quer dizer 'nível' parece essencial para conceber a natureza do pensamento e da consciência. Enquanto os níveis estão isolados uns dos outro, como no Quarto Chinês, tudo é claro, mas se eles interferem e se confundem deixa de o ser. Note-se que o próprio Searle admite que há dois níveis no QC, mas não admite que poderia haver dois pontos de vista (ou abrir-se-ia uma caixa de

Pandora e a experiência e a mente começariam a espalhar-se por toda a parte - de facto Searle pensa que o panpsiquismo é o risco do cognitivismo).

Aqueles que como Dennett e Hofstadter se recusam a admitir que o QC mostre qualquer coisa de importante (i.e., que evidencie a ausência do que quer que seja) defendem que mentes existem em cérebros e podem vir a existir em máquinas programadas. Se e quando essas máquinas vierem a existir os seus poderes causais não derivarão das substâncias de que elas são feitas mas dos programas que instanciam. E poder-se-á saber que elas têm mentes falando com elas e ouvindo com atenção o que têm para dizer...

Em A Redescoberta da Mente Searle apresenta um segundo argumento anti-cognitivista: Esse argumento é, grosseiramente, o seguinte:

A sintaxe não é uma propriedade física
O cognitivismo supõe o tratamento de fenómenos físicos como sintáticos
O cognitivismo incorre na falácia do homúnculo
(i.e., o cognitivismo descreve como se fossem propriedades naturais propriedades que só existem 'para um observador' – no caso, a sintaxe, que não é para Searle uma característica do mundo natural mas uma interpretação por um observador)

As críticas de Searle ao cognitivismo podem ser unificadas notando que para Searle, o princípio da conexão é fundamental para pensarmos na mente. O principio da conexão é o princípio segundo o qual só entendemos como mental o que é actualmente inconsciente porque o entendemos como conteúdo possível da consciência. A consciência é a essência da mente, i.e., só entendemos alguma coisa como mental e não física porque a reportamos à consciência (senão como distinguiriamos um neurónio de uma memória?).

Searle chama à sua posição em teoria da mente um materialismo não reducionista, embora os seus críticos afirmem que não é possivel distinguir esta solução anti-reducionista, segundo a qual os Estados Mentais são causados por e são caracteristicas de, mas não são idênticos a, Estados Cerebrais de um dualismo de propriedades (segundo o qual os fenómenos mentais envolvem propriedades que não são fisicas). No entanto, Searle de modo algum admite ser um dualista.

Searle é anti-reducionista porque defende que a consciência, por ser ontologicamente subjectiva, torna a redução impossivel. Assim, todas as teorias reducionistas do mental falhariam na distinção entre um zombie e um ser consciente. A posição de Searle depende portanto da noção de 'redução'.

A redução é um termo e um problema da filosofia da ciência, um termo para a análise de alguma coisa identificada num nivel de descrição em termos de outro nivel, mais fundamental, permitindo-nos dizer que a 1ª coisa não é mais nada além da 2ª. O problema da redução aplicado ao mental consiste em saber se o mental poderia ser descrito em termos totalmente não mentais. Searle pensa que não, devido à subjectividade ontológica da consciência.

Alguma coisa é 'ontologicamente subjectiva' para Searle se não conseguimos descrever as suas caracteristicas em 3ª pessoa, que é o que estamos a fazer mesmo quando fazemos por exemplo neurofisiologia da consciência. Para Searle, a consciência é uma propriedade física do cerebro apesar da sua subjectividade ontológica e ela é irredutivel a qualquer outra caracteristica fisica. A subjectividade ontológica não deve ser confundida com a subjectividade epistemológica. Uma coisa é a objectividade como a boa tentativa de eliminação das pre-concepções subjectivas, eliminação essa que faz parte do espírito da ciência, outra é a afirmação de que o mundo não contem elementos irredutívelmente

subjectivos. Não há razão para a segunda afirmação. A 1ª noção de objectividade é epistemológica, a 2ª é ontológica.

Ora, se se aceita esta distinção, a verdadeira questão é a seguinte: Como é que podemos ter uma concepção objectiva dos factos ontologicamente subjectivos da consciência? A isto Searle responde com o naturalismo biologico: a ideia de que a consciência é uma caracteristica biológica do cérebro humano e de outros animais e uma propriedade emergente (como a liquidez, a partir da energia cinética molecular).

Searle põe a questão nestes termos porque pensa que muitos materialistas estão errados quando pensam que sem admitir a redução se aceita necessariamente o dualismo.

Os computadores, a filosofia e o funcionalismo

Se através de Dreyfus e Searle se apontou algumas críticas aos fundamentos filosóficos da IA, é preciso por outro lado sublinhar que, não especificamente a IA mas a própria existência de computadores, foi extremamete importante para toda a filosofia da mente dos últimos 40 anos, e central na ideia de funcionalismo.

O funcionalismo é um posição filosófica hoje muito espalhada quanto à natureza da cognição (teorizada por exemplo por H.Putnam). De acordo com o funcionalismo, é essencial para conceber a natureza da cognição a distinção entre hardware e software. I.e., o funcionalismo acentua aquilo a que L. Moniz Pereira (Moniz Pereira 1990) chama a «não obrigatoriedade de correspondência entre o processamento de uma certa função cogntiva e o suporte material que executa esse processamento». Para além de possibiitarem exemplos reais de intencionalidade e racionalidade em sistemas físicos sem necessidade de 'homúnculos' ou observadores cartesianos, os computadores desligaram conceptualmente a ideia de inteligência da particular realização biológica das inteligência nos humanos, e na medida em que a inteligência foi ao longo do tempo considerada definidora daquilo que é ser humano, mudaram a concepção do humano.

Limites?

Actualmente, o debate filosófico em torno da naturalização da cognição, centra-se em dois aparentes limites a uma teoria integralmente científica, feita exclusivamente em 3ª pessoa: os estados qualitativos, i.e., o experienciar, o *what-it's-like-to-be-x* nageliano e a racionalidade, com os seus aspectos normativos.

Sofia Miguens

Bibliografia

BODEN, Margaret (ed), 1990, *Philosphy of AI*, Oxford, Oxford University Press
CHALMERS, David, 1996, *The Conscious Mind*, Oxford, Oxford University Press
DENNETT, Daniel e HOFSTADTER, Douglas (eds.), 1981,*The Mind's I*, New York, Bantam Books
DREYFUS, Hubert, 1972, *What Computers Can't Do*, New York, Harper & Row
MINSKY, Marvin, 1985, *The Society of Mind*, New York, Simon &Schuster
MONIZ PEREIRA, Luís, 2000, Inteligência Artificial, *Intelecto*, 3
NEWELL, Allen, 1990, *Unified Theories of Cognition*, Cambridge MA, Harvard University Press
PINTO, Joâo Alberto, 1999, *Materialismo, Superveniência e Experiência*, Dissertação de Mestrado, Faculdade de Letras da Universidade do Porto

PUTNAM, Hilary, *Minds and Machines*, 1960, in *Philosophical Papers*, vol. 2, Cambridge, Cambridge University Press, 1975
SEARLE, John, 1989, Minds, Brains and Programs, *Behavioral and Brain Sciences*, 3
SEARLE, John, 1998, *A Redescoberta da Mente*, Lisboa, Instituto Piaget
SIMON, Herbert, 1969, *The Sciences of the Artificial*, Cambridge Mass., MIT Press
TURING, Alan, *Computing Machinery and Intelligence* (1950), in DENNETT e HOFSTADTER 1981

Para uma bibliografia relativa às questões filosóficas da IA: 4ª parte da Bibliografia da Filosofia da Mente organizada por David Chalmers, em http://www.u.arizona.edu/~chalmers/biblio.html.