Trans/Form/Ação, São Paulo,
v. 14, p. 109-121, 1991.
https://doi.org/10.1590/S0101-31731991000100006

ROBOTS, INTENCIONALIDADE E INTELIGÊNCIA ARTIFICIAL*

João de Fernandes TEIXEIRA**

RESUMO: O artigo aborda problemas filosóficos relativos à natureza da intencionalidade e da representação mental. A primeira parte apresenta um breve histórico dos problemas, percorrendo rapidamente alguns episódios da filosofia clássica e da filosofia contemporânea. A segunda parte examina o Chinese Room Argument *(Argumento do Quarto do Chinês) formulado por J. Searle. A terceira parte desenvolve alguns argumentos visando mostrar a inadequação do modelo funcionalista de mente na construção de* robots. *A conclusão (quarta parte) aponta algumas alternativas ao modelo funcionalista tradicional, como, por exemplo, o conexionismo.*

UNITERMOS: Intencionalidade; representação mental; inteligência artificial; funcionalismo; robots*; conexionismo.*

I

A questão da natureza da intencionalidade e da representação mental constitui um dos mais tradicionais problemas da Filosofia. Em diversos episódios da história das idéias, os filósofos tentaram responder a pergunta de como é possível para nossa mente formar uma representação do mundo exterior que servisse de guia não só para sua cognição como, também, para a orientação de nossos próprios comportamentos.

O surgimento da filosofia cartesiana, na época clássica, constitui um marco importante na história desta questão. Com efeito, deve-se a Descartes a revolução filosófica que trouxe como conseqüência fundamental a separação entre mente e corpo, ou entre *res cogitans* e *res extensa*. Mas ao mesmo tempo em que efetuava esta mudança paradigmática na história da Filosofia, o cartesianismo deixava como herança não apenas o problema de formular uma articulação entre as duas substâncias de modo a poder explicar o comportamento, como também a questão da própria possibilidade de cognição do chamado mundo exterior. Uma vez postulada a existência de duas substâncias distintas e sua mútua irredutibilidade, restava saber como nossas idéias (ou representações) poderiam espelhar esse mundo fora de nós e para além dos limites de nossa mente.

* Comunicação apresentada na XIV Jornada de Filosofia e Ciências Humanas – UNESP – Marília – outubro de 1990.

** Departamento de Filosofia de Faculdade de Filosofia e Ciências – UNESP – 17500 – Marília – SP.

Uma solução parcial para o problema, oferecida pelo cartesianismo, foi sustentar a existência de um Deus Todo-Poderoso e não enganador que asseguraria a correspondência entre o mundo e a representação mental que dele formamos. Este Deus – o famoso *deus ex machina* cartesiano – constituiria uma espécie de garantia metafísica de que o conhecimento e a representação correta do mundo exterior poderiam ser alcançados.

Contudo, ao relermos a obra de Descartes, notamos hoje que a idéia de um *deus ex machina* parece muito mais uma hipótese *ad hoc* do que uma real tentativa de resolver o problema da natureza da representação mental. E alguns anos mais tarde, num episódio posterior da história da Filosofia, a questão parece ser retomada, desta vez com a obra de I. Kant. Com efeito, se tivéssemos de fazer um resumo brutal e ao mesmo tempo grosseiro do conteúdo da primeira Crítica kantiana, poderíamos dizer que Kant buscava investigar como que operações internas de nossa mente poderiam captar relações objetivas estabelecidas no mundo externo, tornando assim possíveis o conhecimento e a representação adequada de uma realidade independente de nós. A solução proposta por Kant – muito mais interessante do que a cartesiana – foi afirmar que a realidade que enxergamos e representamos é moldada pelas operações do nosso aparelho cognitivo. Assim, o estudo deste último proporcionaria ao mesmo tempo o conhecimento da única versão do mundo que nos é disponível – uma versão que não é, contudo, subjetiva, mas limitada pelo modo como nossas representações podem se articular.

Os séculos XIX e XX contribuíram com mais alguns episódios interessantes na história do problema da natureza da representação mental. Em 1925, o psicólogo e filósofo Franz Brentano publica a *Psychologie von empirischen Standpunkt* e, numa de suas passagens freqüentemente citadas afirma que:

> *Every mental phenomenon is characterized by what the scholastics of the Middle Ages called the intentional (and also mental) inexistence of an object, and what we could call, although in not entirely unambiguous terms, the reference to a content, a direction upon an object (by which we are not to understand a reality in this cases) or an immanent objectivity* (1, p. 50).

Brentano reviveu o termo *intencionalidade* de textos de filósofos medievais. Ele não queria pura e simplesmente reviver o termo resgatando seu significado original, mas também explorar algumas das conseqüências advindas de sua re-adoção. Uma delas (e cremos que seja a mais importante) foi asseverar a existência de um hiato entre o físico e o mental, derivado da impossibilidade de sistemas físicos virem a gerar estados intencionais.

Neste ponto Brentano parece ter sido bastante fiel à tradição medieval que o inspirara. Os filósofos medievais julgavam que, todas as vezes que pensamos, por exemplo, numa rosa, uma rosa entra na nossa mente. Claro que a rosa que entra na nossa mente não pode ser uma rosa no sentido *físico*, e daí eles julgarem que a geração de idéias e representações requer a capacidade de nossa mente receber formas imateriais – algo que só seria possível na medida em que nossa mente nada tem de físico.

A imaterialidade da mente seria, assim, um pré-requisito da sua própria faculdade de representar objetos do mundo exterior.

Mas a importância histórica das teses de Brentano não se resume a re-afirmação da imaterialidade do mental. É a partir de seus textos que se inicia uma aproximação entre as questões envolvendo a natureza da representação mental e aquelas envolvendo a natureza da intencionalidade e dos estados intencionais, que passam a ser vistas como constituindo um único problema filosófico amplo e geral. Assim sendo, representar significa estabelecer uma relação entre um conteúdo mental e o possível objeto a ele correspondente no mundo exterior – uma relação que pressupõe uma direcionalidade em relação a este objeto. Determinar a natureza desta relação e da direcionalidade que a torna possível significa encontrar uma solução para o problema da intencionalidade e da natureza da representação mental.

Ecos desta nova maneira de formular a questão tradicional da representação mental podem ser encontrados na literatura contemporânea sobre Filosofia da Mente, como por exemplo a seguinte passagem da *Análise da Mente* de B. Russell:

> "Vamos supor que estamos pensando na Catedral de S. Paulo. (...) temos de distinguir três elementos que necessariamente se combinam na constituição desse pensamento. Primeiro, temos o ato de pensar, que será o mesmo, independentemente do que estivermos pensando. Depois temos o que caracteriza o pensamento em comparação com outros pensamentos: é o conteúdo. E finalmente, temos a Catedral de S. Paulo, que é o objeto de nosso pensamento. Deve haver uma diferença entre o conteúdo de um pensamento e aquilo sobre o que o pensamento recai, dado que o pensamento está aqui, ao passo que aquilo sobre o que o pensamento recai pode não estar; portanto vê-se claramente que o pensamento não é o mesmo que a Catedral de S. Paulo. Isto parece mostrar que devemos distinguir entre conteúdo e objeto. (...) O objeto pode existir sem o pensamento, mas o pensamento não pode existir sem o objeto: os três elementos, ato, conteúdo e objeto são todos obrigados a constituir uma única ocorrência chamada "pensar na Catedral de S. Paulo." (3, p. 14)

Se lermos esta passagem de B. Russell com a devida atenção, verificaremos que nela se esboça nitidamente a relação conteúdo-objeto do pensamento bem como a idéia de uma direcionalidade (intencionalidade) presente na relação que dá lugar a representação mental.

II

O surgimento de disciplinas novas tais como a Inteligência Artificial e a Ciência Cognitiva que presenciamos nas últimas décadas, reavivou os debates filosóficos acerca da natureza da representação mental. A Inteligência Artificial, na tentativa de realizar um de seus principais projetos, qual seja, a simulação do comportamento inteligente, deparou-se com a questão: Pode um sistema artificial gerar estados inten-

cionais e representar o mundo exterior? Será que, como supunha Brentano, a intencionalidade determina um hiato intransponível entre o físico e o mental de tal maneira que a geração de estados intencionais através da construção de sistemas artificiais (físicos) estaria inevitavelmente condenada ao fracasso? Esta indagação traz uma nova dimensão para o problema filosófico tradicional da representação mental; uma dimensão que extravasa os limites de uma polêmica puramente filosófica. Afinal, da capacidade de gerar estados intencionais depende a própria possibilidade de simular comportamentos inteligentes, dotados de elevado grau de flexibilidade, uma vez que a representação do mundo exterior contribui diretamente para sua adaptação face a mudanças nas circunstâncias ambientais.

Este debate toma um impulso decisivo a partir da publicação do artigo de J. Searle, *Minds, Brains and Programs* em 1980. Neste artigo, Searle apresenta uma crítica vigorosa da possibilidade de se obter um equivalente mecânico para o fenômeno cognitivo humano que normalmente denominamos *compreensão*. Seu ponto de partida é a análise dos programas para compreender histórias curtas desenvolvidas por R. Schank na Universidade de Yale e que se concretizaram no trabalho *Scripts, Plans, Goals and Understanding*.

Os programas desenvolvidos por R. Schank – conforme assinalamos – têm por objetivo a compreensão de histórias. Por exemplo, se se fornece a um computador o seguinte relato: um homem entra num restaurante, pede um sanduíche e sai sem pagar ou deixar gorjeta porque notou que o pão estava estragado, o programa de Schank é construído de tal maneira que lhe é possível responder coerentemente questões elaboradas com base no texto da história. Tendo em vista estes resultados, Schank sustenta que este tipo de programa é capaz de *compreender* o texto e constituir uma explicação para a capacidade do ser humano de compreender textos ou histórias curtas.

As críticas desenvolvidas por Searle às pretensões de que um tal programa realmente compreende baseiam-se na construção de um experimento mental que reproduz o procedimento do próprio programa. O caminho adotado por Searle para construir este experimento mental é o inverso do procedimento normalmente utilizado para elaborar simulações cognitivas: trata-se de instanciar o programa de Schank num sujeito humano.

Imagine um falante trancado num quarto. Este falante só conhece o Português e tem em seu poder: a) um texto escrito em Chinês; b) um conjunto de regras de transformação, em Português, que permite executar operações sobre o texto em Chinês. Estas operações são idênticas àquelas desempenhadas pelos programas de Schank: trata-se de operações de decomposição e recomposição de palavras com base num *script* que contém informação relevante (por exemplo, sobre como são os restaurantes em geral, qual o procedimento para pedir comida, etc.) o que capacita o sistema a responder as questões desejadas.

O falante (trancado no quarto) recebe periodicamente novos textos em Chinês e aplica estas operações ou regras de transformação associando as seqüências anteriores com as seqüências mais recentes. Com base nestas regras de transformação ele passa

a emitir ou escrever mais seqüências de símbolos em Chinês. Claro que o falante preso no quarto não sabe precisamente o conteúdo das informações que ele está gerando com base nos dois textos e nas regras de transformação. O primeiro texto corresponde, no nosso experimento mental, ao relato que é elaborado com base neste relato, e, as novas seqüências geradas, às respostas a estas questões. As regras de transformação são bastante complexas, e concebidas de maneira tal que elas simulem os processos mentais e o comportamento lingüístico de um falante nativo de Chinês numa conversação habitual. Após um certo tempo o falante aprendeu a manipular perfeitamente estas regras de transformação, e, com base nos *outputs* um observador externo poderia dizer que ele *compreende* Chinês – o que, no entender de Searle constitui um contra-senso.

A instanciação dos programas de Schank num sujeito humano, reproduzida neste experimento mental, é, para Searle, bastante reveladora. Ela mostra que os programas desse tipo não estabelecem as condições necessárias para a simulação da atividade cognitiva da *compreensão*: o falante aplica as regras de transformação e compreende estas regras, mas as seqüências de símbolos em Chinês não têm nenhum significado para ele. A manipulação de símbolos realizada no programa é inteiramente cega.

Ademais, como ressalta Searle na resposta às objeções ao seu texto *Minds, Brains and Programs*, "a manipulação de símbolos formais por si só não tem intencionalidade, não é sequer manipulação de símbolo, uma vez que esses símbolos não simbolizam nada. Eles têm apenas sintaxe, mas não semântica". (4, p. 300)

Ainda na sua resposta às objeções, Searle ressalta que a esses programas "falta aquilo que chamarei de intencionalidade intrínseca ou de genuínos estados mentais". (4, p. 305)

A atribuição de intencionalidade ou de *significado* diz Searle, é, nestes casos, sempre uma atribuição *a posteriori*, dependente da intencionalidade intrínseca dos sujeitos humanos que observam os *outputs* do programa.

Mas o que é "intencionalidade intrínseca" no entender de Searle, e que parentesco tem esta noção com a idéia de significado? O conceito de intencionalidade intrínseca não é largamente explorado em *Minds, Brains and Programs*. Searle o desenvolve com maior profundidade em outros artigos, tais como *Intrinsic Intentionality* e *What are Intentional States* (1982), bem como no seu livro de 1983, *Intentionality*. A intencionalidade, segundo Searle, é uma "capacidade" apresentada pelos seres vivos, através da qual nossos estados mentais se relacionam com os objetos e estados de coisas no mundo. Assim, se tenho uma intenção, esta intenção deve ser a intenção de fazer alguma coisa, se tenho um desejo ou um medo, este desejo e este medo devem ser um desejo ou medo de alguma coisa que está no mundo. Um estado intencional pode ser definido, grosso modo, como uma representação associada a um determinado estado psicológico.

Esta mesma capacidade – estritamente biológica no entender de Searle – percorre nossa linguagem, convertendo-a num tipo particular de relação organismo/mundo. Contudo, ela não é uma propriedade da linguagem e sim uma propriedade específica

que nossos estados mentais imprimem ao nosso discurso. Nesta operação, os sinais lingüísticos, sejam eles os sons que emitimos ou as marcas que fazemos no papel, passam a ser representações de coisas ou estados de coisas que ocorrem no mundo, e no caso específico das representações lingüísticas podemos afirmar que elas constituem descrições dessas representações ou mesmo representações de representações que estão na nossa mente. A intencionalidade dos estados mentais não é derivada de formas mais primárias da intencionalidade, mas é algo intrínseco aos próprios estados mentais. Neste sentido, a intencionalidade é a propriedade constitutiva do mental e sua base é estritamente biológica – só os organismos desempenham esta atividade relacional com o mundo, constituindo representações. Sua origem está nas próprias operações do cérebro e na sua estrutura, constituindo parte do sistema biológico humano, assim como a circulação do sangue e a digestão.

A intencionalidade intrínseca, presente no discurso lingüístico, constitui uma forma derivada de intencionalidade que consiste na relação das representações lingüísticas com os estados mentais intencionais, o que permite que estas últimas sejam representações de alguma coisa do meio ambiente. Em outras palavras, esta relação entre representações lingüísticas e estados intencionais transforma o código lingüístico num conjunto de signos, ou seja, estabelece o seu *significado*. Neste sentido, a intencionalidade intrínseca constitui para Searle a condição necessária para que um sistema simbólico adquira uma dimensão *semântica*. Sem esta dimensão semântica não podemos falar de compreensão. E sem esta relação entre representações mentais ou conteúdos intencionais e representações lingüísticas não podemos falar de compreensão de textos ou compreensão lingüística.

A ausência da intencionalidade intrínseca nos programas desenvolvidos por Schank está na base da afirmação de Searle de que estes últimos constituem um procedimento cego de associação de signos sem significado – um procedimento cego que não deve ser confundido com autêntica compreensão lingüística.

Ora, até que ponto podemos supor que as afirmações de Searle são corretas? Se o forem, a questão que formulamos no início desta secção estaria respondida em caráter definitivo, ou seja, sistemas artificiais não podem gerar estados intencionais e nem tampouco representar o mundo exterior.

Ocorre que vários filósofos favoráveis ao projeto da Inteligência Artificial apresentaram vários contra-argumentos às posições defendidas por Searle. Alguns deles apontaram falhas no argumento principal, salientando que não sabemos se de fato os computadores podem ou não compreender alguma coisa. A situação seria semelhante a quando observamos um ser humano responder a perguntas acerca de um texto qualquer: como podemos estar certos de que essa pessoa compreende o que está fazendo? Por acaso muitos de nossos processos mentais cotidianos não são tão rotineiros que os fazemos por uma associação tão mecânica e cega como as do computador? Se as operações efetuadas pelo falante trancado no quarto são cegas, será que não podemos afirmar o mesmo de nossas próprias operações mentais? Mesmo quando tentamos examinar o fluxo de nossos pensamentos, isto não nos dá nenhuma informação acerca

de como ocorrem as operações de nosso cérebro. Somos, em grande parte, opacos para nós mesmos – e não seria essa uma situação idêntica àquela de alguém que olha para os resultados das operações de um computador e, com base nestes últimos, quer sustentar a afirmação de que essa máquina nada compreende acerca dessas operações?

É difícil saber quem tem razão num debate deste tipo: como todas as polêmicas filosóficas, esta também deve ser inconclusiva. Entretanto, é preciso fazer uma observação importante: a crítica de Searle pode ser considerada correta se levarmos em conta o tipo de modelo computacional da mente sobre a qual ela recai. Trata-se de um modelo muito específico e que vigorou até meados dos anos 80, qual seja, o *funcionalismo.*

O funcionalismo baseia-se na tese de que a essência de nossos estados psicológicos reside na sua interconexão (às vezes causal) com outros estados, formando uma complexa economia de estados internos que media os *inputs* do meio ambiente e os *outputs* comportamentais. A natureza e a razão de ser destes nossos estados mentais – sejam eles crenças, volições, imagens mentais ou o que quer que seja – são determinadas *funcionalmente* tendo em vista os objetivos e tarefas que um sistema vai realizar. A peculiaridade desta concepção reside, contudo, no fato de que os estados mentais e sua natureza são definidos e constituídos através de um conjunto de relações abstratas que eles mantêm entre si. Este conjunto pode ser instanciado em diferentes mecanismos ou sistemas, sejam eles computadores antigos com suas válvulas, sejam organismos com sua estrutura biológica como é o caso do *homo sapiens.* Se o sistema é capaz de instanciar este conjunto de relações abstratas, ele será um sistema mental ou uma "mente".

A contrapartida computacional desta concepção de mente será a idéia de que a produção de comportamento inteligente baseia-se na manipulação adequada de símbolos, de acordo com um conjunto de regras que uma vez aplicadas a estes símbolos devem gerar procedimentos efetivos (Nota A). A mente é uma máquina que realiza computações, isto é, que opera sintaticamente sobre representações. Uma determinada tarefa é computável se ela puder ser concebida de tal forma que seja reduzida a um conjunto de representações e de instruções (regras, programas) que opere sobre elas.

Ora, até que ponto este modelo de mente será válido? Discutir a validade do argumento de Searle parece pressupor, num primeiro momento, uma discussão dos limites do modelo funcionalista.

III

Numa passagem bastante significativa, o lingüista e filósofo norte-americano J. Fodor procura definir a noção de computação que norteia a elaboração dos modelos funcionalistas da mente:

> "*I take it that computational processes are both* symbolic and formal. *They are defined over representations, and they are formal because they apply to representations in virtue of (roughly) the* syntax *of representations.*" (2, p. 226)

Ora, haverá um limite para nossa capacidade de transformar todas as nossas representações do mundo em representações do tipo simbólico? E até que ponto sistemas artificiais, construídos com base no modelo funcionalista de mente poderão efetivamente vir a gerar representações do mundo externo?

Podemos iniciar, esta discussão, imaginando uma situação na qual desejemos construir um *robot*, com a finalidade de executar uma ampla variedade de tarefas num meio ambiente real. Para usarmos uma metáfora familiar, podemos supor que este *robot* tem uma forma semelhante a de um ser humano, com a CPU (*Central Processing Unit*) correspondendo ao cérebro (ou mente!). A CPU controla o movimento dos periféricos que, dotados de censores, transmitem dados acerca do meio ambiente. Estes dados são, por sua vez, transformados e processados pela CPU que os utiliza para orientar os comportamentos do *robot*. Em outras palavras, a CPU deste *robot* elabora os *inputs* que chegam, transformando-os em representações do seu meio ambiente, que passam a desempenhar o papel de estados internos que vem se somar a outros estados internos que possam porventura serem gerados pelas suas próprias operações. Isto torna nossa máquina imaginária bastante sofisticada, dotada de um elevado grau de autonomia e de complexidade que se aproximam daquela de um ser humano, possibilitando a auto-organização de seu próprio repertório de comportamentos.

Ora, tudo se passa como se a CPU deste *robot* fosse uma espécie de "câmara cega", onde ocorre um enorme fluxo de estados internos. A situação é semelhante àquela do nosso cérebro, que é um palco por onde passa um enorme fluxo de informações, sejam estas provenientes das próprias atividades cerebrais e orgânicas, ou provenientes do mundo externo através dos sentidos. Ocorre, porém, que o cérebro não pode "ver" de onde provém as informações (ele é uma "câmara escura") e tem de decidir, unicamente com base no seu conteúdo, quais são os estados internos legitimamente representacionais e quais são aqueles que resultam de sua própria atividade. O mesmo ocorre na CPU de nosso *robot*, que teria de realizar esta tarefa: distinguir entre estados representacionais e outros estados internos sem o auxílio de uma programação prévia.

A dificuldade, de estabelecer a distinção entre estados internos de natureza representacional e outros tipos de estados internos, pode se converter num verdadeiro obstáculo para a possibilidade de atribuir estados intencionais a sistemas artificiais. Dois casos típicos ilustram esta dificuldade:

a) *O problema da memória* – Considere-se novamente o caso do nosso *robot*. Trata-se de uma réplica mecânica de um ser humano adulto. Ele não dispõe de um passado ou uma "infância". Podemos suprir esta deficiência através de um "implante de memória", ou seja, o implante de uma memória artificial, composta de proposições (lembranças de proposições, em geral expressas verbalmente, tornando-se memória auditiva) ou mesmo como um conjunto de imagens mentais caprichosamente elaboradas e implantadas. Uma vez implantado este dispositivo, nossa réplica mecânica, provavelmente, apresentará um comportamento externo *como se tivesse uma memória*, sendo capaz de responder adequadamente a questões acerca de seu suposto

passado, numa cena algo familiar de ficção científica que encontramos no filme *Blade Runner*.

Entretanto, a similaridade de comportamento exibida pela réplica será apenas aparente: sem programação prévia não há como distinguir dentre estados internos aqueles que correspondem a memórias e aqueles que correspondem a estados mentais referentes a eventos no presente. Se as "memórias" forem expressas na forma de conteúdos mentais de caráter imagético, elas se tornam indistinguíveis de outros conteúdos imagéticos que normalmente classificaríamos como imagens mentais, imaginação, etc. No caso das proposições a situação não é diferente: exprimir uma proposição como ocorrendo no passado, através de uma sentença com um verbo no passado não significa referir-se a algo que ocorreu anteriormente no tempo. No caso de um ser humano a dificuldade poderia ser suplantada: conteúdos mentais relacionados a lembranças seriam distintos de outros na medida em que incorporariam e conservariam (parcialmente) uma relação causal com os eventos passados que efetivamente os produziram.

b) *O problema da percepção e da alucinação* – Se construirmos um *robot* e nele adaptarmos um mecanismo de visão artificial que sirva para orientar suas ações no meio ambiente, é bastante provável que tal mecanismo de visão artificial não gere percepções e sim alucinações verídicas. Entendemos por alucinação verídica um tipo peculiar de experiência visual no qual o conteúdo imagético e informacional coincide com aquele de uma percepção genuína. Por exemplo, teremos um caso de alucinação verídica quando uma pessoa tiver diante de si uma paisagem e simultaneamente aluciná-la com todos os seus detalhes. A cena real e a experiência visual (alucinatória) coincidem perfeitamente em termos de conteúdo imagético e informacional, embora essa pessoa possa até mesmo estar vendada, e suas experiências visuais estarem sendo produzidas por uma neurocirurgia.

Pode um sistema artificial distinguir percepções de alucinações verídicas? Tomemos como ponto de partida um sistema que consiste de uma CPU conectada a um "olho mecânico" – algo parecido com uma câmara de TV, responsável pela produção de imagens do meio ambiente onde este sistema se move. Supondo que este mecanismo de visão artificial sirva para a orientação do comportamento, parece haver, neste caso, apenas duas possibilidades. A primeira consistiria em estabelecer todas as ações – e, conseqüentemente, todas as "percepções" ou "experiências visuais" que seriam produzidas pelo seu olho mecânico. Assim, por exemplo, poder-se-ia programar o *robot* para andar do centro de Paris até a Torre Eiffel, ao mesmo tempo que seu mecanismo de visão artificial produzisse experiências visuais com um conteúdo inteiramente semelhante àquelas que um ser humano tem quando caminha do centro de Paris até a Torre Eiffel. A situação, neste caso, não difere muito daquela na qual o neurofisiólogo introduz agulhas e eletrodos no cérebro de seu paciente, produzindo imagens visuais. Este papel é desempenhado pelo programador que, além de produzir alucinações verídicas na sua máquina, estaria também controlando suas ações. O grau de autonomia desta máquina seria mínimo e mesmo que, por hipótese, seu dispositivo de visão fosse capaz de gerar percepções, tal dispositivo tornar-se-ia praticamente dispensável.

A segunda alternativa consistiria em projetar um *robot* com alto grau de autonomia, dotado apenas de algumas metas internas gerais. Esta máquina mais sofisticada e não inteiramente pré-programada teria um mecanismo de visão que se encarregaria de sua interação com o meio ambiente, possibilitando a auto-organização do seu próprio repertório de comportamentos. Mas como funcionaria este mecanismo de visão? Se se tratar de algo parecido com a câmara de TV de que falamos a pouco, também, neste caso, este dispositivo estaria muito mais próximo da geração de alucinações verídicas do que de percepções. A relação entre as imagens registradas pela câmara e as cenas do mundo a elas correspondentes seria estabelecida através de uma relação causal; como poderia tal máquina distinguir, dentre seus estados internos de caráter imagético quais seriam aqueles que correspondem às suas percepções e a partir destas produzir comportamentos adequados?

Se projetarmos uma máquina altamente sofisticada, ela será dotada de capacidade de gerar estados internos de caráter imagético: memórias, imagens mentais, etc. Se esta máquina não tiver sido previamente programada para tal, ela não terá condições de gerar uma distinção entre processos internos e processos externos. A incapacidade de estabelecer esta distinção sem o auxílio de programação prévia faz com que suas percepções sejam assimiláveis ao fluxo de processos internos, tornando-as indistinguíveis das alucinações verídicas que poderiam igualmente estar ocorrendo no seu interior. Esta máquina, por mais sofisticada que fosse, não estaria percebendo o mundo a sua volta: ela não pode *representá-lo*. Seu sucesso na execução de tarefas e na produção de comportamentos adequados ao seu meio ambiente poderia ser muito grande, mas tratar-se-ia, no melhor dos casos, de uma adequação cega, que não pressupõe a geração de estados intencionais. Seria uma situação estruturalmente semelhante ao produtor de textos em Chinês de que nos fala J. Searle (Nota B).

Os dois problemas de que tratamos acima parecem sugerir uma resposta negativa no que diz respeito a possibilidade de sistemas artificiais virem a gerar estados intencionais. Eles apontam para a existência de grandes obstáculos que, ainda precisam ser superados para que isto seja possível. É preciso, agora, tentar responder a nossa primeira indagação, qual seja, se haverá um limite para nossa capacidade de transformar todas as nossas representações do mundo em representações do tipo simbólico. Abordaremos esta questão examinando um terceiro conjunto de problemas.

c) *O problema da auto-referência* – Nosso ponto de partida será novamente um *robot* com alto grau de autonomia comportamental e semelhança em relação a um ser humano. Poderia este *robot*, em alguma ocasião, ter o pensamento "estou aqui"?

Na literatura contemporânea sobre Filosofia da Mente desenvolveu-se a visão de que a base hierárquica das representações possíveis é constituída por conteúdos mentais pré-representacionais ou pré-conceituais. Esta concepção fundamenta-se na idéia de que o caráter significativo que estruturas mentais possam adquirir não depende apenas de sua manifestação local no organismo (ou no cérebro), mas da totalidade do ambiente onde esse organismo atua. Assim, a determinação do caráter significativo de certos conteúdos mentais não seria dada unicamente por características intrínsecas

que estes pensamentos possam exibir, mas pela sua ocorrência contextual estabelecida numa relação entre organismo e meio ambiente. Estes estados mentais – cujo caráter intencional é estabelecido por esta relação simbiótica com o meio ambiente – são chamados pensamentos *de re* e constituem a base hierárquica das representações possíveis. Seu traço distintivo consiste no fato de eles implicitamente incorporarem um vínculo com a existência de seus referentes no mundo – um vínculo que é responsável não apenas pela própria possibilidade de sua ocorrência na vida mental do organismo, como também pela sua própria identidade enquanto pensamentos.

A natureza intencional (a direcionalidade) deste tipo de pensamentos básicos é dada por fatores contextuais que compõem circunstâncias que participam da ocorrência destes pensamentos, e os dotam de uma dimensão semântica oriunda da relação do sujeito (ou organismo) com uma realidade extramental. Assim, por exemplo, os partidários deste tipo de visão, acerca da natureza de pensamentos básicos (*de re*), têm chamado atenção para a necessidade de se distinguir entre ter um pensamento *sobre* Londres de um pensamento *acerca de* Londres. Pois para ter um pensamento *sobre* Londres eu posso estar em qualquer lugar, bastando apenas que minha mente forme algumas proposições ou imagens das torres do Parlamento e da *Oxford Street*, enquanto para ter um pensamento *acerca de* Londres (e se este pensamento deve também dar minha localização espacial), é preciso que eu (meu organismo) esteja fisicamente em Londres como condição necessária. A distinção não é puramente lingüística como poderíamos ser levados a crer: ela aponta para uma divisão fundamental entre dois *tipos* de pensamentos, pois o pensamento *acerca de* Londres (*de re*) tem sua dimensão semântica estabelecida pelas circunstâncias contextuais em que ocorre o pensamento, enquanto o pensamento *sobre* Londres independe destas últimas (e constitui um pensamento *de dicto*).

Substanciando este ponto de vista poder-se-ia afirmar a impossibilidade de se construir um *robot* ou um mecanismo que fosse capaz de ter o pensamento "estou aqui". O problema é que este pensamento específico (*de re*), "estou aqui", teria de ser *representado no programa central que controla o robot*, mas aqui encontramos uma dificuldade fundamental: se o caráter semântico deste tipo de pensamento depende do fato de ele incorporar um fator extramental derivado de uma ligação com as circunstâncias de sua ocorrência contextual, o pensamento do tipo "estou aqui" não adquire significado pelo fato de eventualmente constituir-se como uma representação. Isto porque o significado de pensamentos deste tipo não decorre de nenhum tipo de conteúdo especificável, seja ele imagético ou proposicional: mesmo que em algum momento do programa do *robot* aparecesse a proposição "estou aqui", ela poderia nada dizer acerca da posição do *robot* no seu meio ambiente. Em outras palavras, o caráter transcendente (e semântico) de um pensamento *de re* resiste a sua transformação plena numa representação ou num pensamento *de dicto* plenamente conceitualizado. Mas haverá outra maneira de programar a CPU do *robot* sem antes conceitualizar todas as operações que esta deve realizar? (Nota C)

IV

Os problemas que delineamos na secção anterior não nos obrigam a validar a conclusão de que o projeto científico da Inteligência Artificial está necessariamente fadado ao fracasso. Esta é a conclusão que pensadores com J. Searle gostariam de extrair das proposições acima, para tentar esboçar mais um argumento em favor da impossibilidade de simulação do comportamento inteligente. Mas esta não é a conclusão que desejamos extrair, sobretudo na medida em que não partilhamos dos pontos de vista de J. Searle.

Não nos parece que os objetivos propostos pela Inteligência Artificial sejam inatingíveis. Quem sustenta esta perspectiva corre o mesmo tipo de risco que correram aqueles que, no século passado, sustentaram a idéia de que o homem jamais poderia chegar a lua. Ocorre, entretanto, que o modelo computacional de mente, que tem sido utilizado por muitos teóricos da Inteligência Artificial, precisa de uma ampla revisão. Esta revisão, que já se iniciou a alguns anos atrás, implica uma reformulação conceitual de grande porte que visa redefinir a metáfora a ser empregada na descrição dos fenômenos mentais. A elaboração de uma nova metáfora implica o progressivo abandono de noções pioneiras da Inteligência Artificial, como por exemplo a própria máquina de Turing e os pressupostos básicos do funcionalismo tradicional. Trata-se de uma tarefa que já se encontra em marcha, protagonizada pelo aparecimento dos chamados modelos conexionistas.

Estes novos modelos reaproximam a Computação da Neurofisiologia e da Biologia Evolucionária, procurando desenvolver uma nova metáfora onde, por exemplo, o conhecimento pré-proposicional ou subsimbólico poderá ser adequadamente acomodado, resolvendo algumas das dificuldades que a Inteligência Artificial tradicional não pôde superar. "O cérebro é a melhor metáfora para falarmos da mente" – eis o que parece sugerir o conexionismo, num empreendimento teórico de grande envergadura, que visa à integração de vários tipos de conhecimento que até então permaneceram estanques, isto é, sem uma articulação interdisciplinar. Do conexionismo e da profunda revolução teórica que nele está contida, teremos oportunidade de falar em outro trabalho.

NOTAS

A – Para se ter uma noção mais detalhada de Máquina de Turing e de procedimento efetivo ver: Teixeira J. de F. *O que é Inteligência Artificial*, cap. 2.

B – Um estudo mais detalhado dos problemas envolvidos na percepção e na elaboração de sistemas artificiais de visão está no artigo de minha autoria "A Máquina de Enxergar" (Revista Discurso, n. 19, 1991) – Departamento de Filosofia da USP. (No prelo). Trecho semelhante ao explorado, neste item, encontra-se no referido artigo.

C – Para um estudo da oposição *de re* e *de dicto* ver o artigo de minha autoria "Inteligência Artificial e Caça aos Andróides" (Psicologia, 1990, número especial sobre Filosofia da Psicologia, no prelo). Trecho semelhante ao explorado, neste item, encontra-se no referido artigo.

TEIXEIRA, J. de F. Robots, intentionality and artificial intelligence. **Trans/Form/Ação,** São Paulo, v. 14, p. 109-121, 1991.

ABSTRACT: The paper focuses in philosophical problems concerning the nature of intentionality and mental representation. The first part presents a historical outline of the problem and reviews some classical/contemporary writings on the question. The second part examines the so-called Chinese Room Argument formulated by J. Searle. The third part presents a few arguments aiming to show the inadequacy of the functionalist model for the design of robots. The conclusion points to some alternatives to the traditional functionalist model such as, for instance, the connectionist model.

KEYWORDS: Intentionality; mental representation; artificial intelligence; functionalism; robots; connectionism.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BRENTANO, F. *Psychology from an empirical standpoint.* Trad. A. C. Pancurello, D. B. Terrell and L. L. McAlister. New York: Humanities Press, 1925/1973.
2. FODOR, J. *Representations:* philosophical essays on the foundations of cognitive science. Cambridge, MA: The MIT Press, 1981.
3. RUSSEL, B. *A análise da mente.* Trad. Antonio Cirurgião. Rio de Janeiro: Zahar, 1971/1976.
4. SEARLE, J. Minds, brains and programs. In: HAUGELAND, J., ed. – *Mind design.* Vermont: Bradford Books, 1981.

BIBLIOGRAFIA CONSULTADA

DESCARTES, R. *Meditações.* Trad. J. Guinsburg e Bento Prado Jr. São Paulo: Abril Cultural, 1979. (Os pensadores).

KANT, I. *Crítica da razão pura.* Trad. Valério Rohden e Udo B. Moosburger. São Paulo: Abril Cultural, 1980. (Os pensadores).

SCHANK, R. *Scripts, plans, goals and understanding.* Hillsdale: N. J. Lawrence Erlbaum, 1977.

SEARLE, J. *Intentionality.* Cambridge: Cambridge University Press, 1983.

SEARLE, J. Intrinsic intentionality. *Behavioural and Brain Science,* Cambridge, v. 3, 1980.

SEARLE, J. What are intentional states. In: DREYFUS, H., ed. – *Husserl, intentionality and cognitive science.* Vermont: Bradford Books, 1982.

TEIXEIRA, J. de F. *O que é inteligência artificial.* São Paulo: Brasiliense, 1990. (Primeiros Passos, nº 230)