UNIVERSIDADE FEDERAL DE GOIÁS

DEPARTAMENTO DE FILOSOFIA

MESTRADO EM FILOSOFIA

INTENCIONALIDADE E INTELIGÊNCIA ARTIFICIAL NO PENSAMENTO DE DENNETT

GUILHERME SILVEIRA DE ALMEIDA SANTOS

GOIÂNIA

2013

UNIVERSIDADE FEDERAL DE GOIÁS

DEPARTAMENTO DE FILOSOFIA

MESTRADO EM FILOSOFIA

INTENCIONALIDADE E INTELIGÊNCIA ARTIFICIAL NO PENSAMENTO DE DENNETT

GUILHERME SILVEIRA DE ALMEIDA SANTOS

ORIENTADOR: Prof. Dr. ANDRÉ PORTO

GOIÂNIA

2013

Agradecimentos,

ao Prof. Dr. André Porto,

pela orientação deste trabalho.

“ tanto o filósofo quanto o cientista estão no mesmo barco”

W. V. Quine

SUMÁRIO

RESUMO

SANTOS, Guilherme Silveira de Almeida. *Intencionalidade e Inteligência Artificial*. Universidade federal de Goiás. Goiânia, 2010.

O conceito de postura intencional, um conceito de suma importância no pensamento filosófico de Daniel Dennett, é um aspecto central para um ponto de vista naturalizado da intencionalidade.

Há três modos distintos para predizer o comportamento de um objeto:

1) A postura física, através da qual a predição é feita baseando-se nas propriedades físicas ou leis físicas

2) A postura de projeto, onde consideramos a função de um objeto.

3) A postura intencional, através da qual consideramos um objeto como um agente intencional, dotado de crenças, pensamentos e intenções.

Adicionalmente, do ponto de vista de Dennett,em parte, a naturalização da intencionalidade é o caminho para a possibilidade de construção de computadores que apresentarão comportamento intencional. Comparativamente, o desenvolvimento de computadores inteligentes é o objetivo da pesquisa de inteligência artificial (IA).

Ademais, é importante mostrar as refutações de dois argumentos céticos que tentam provar a impossibilidade da intencionalidade em máquinas. Dois argumentos críticos a certos aspectos da pesquisa de IA são o argumento do teorema de Gödel e o argumento do quarto chinês.

O objetivo da dissertação é mostrar que o conceito de postura intencional é uma possibilidade para a construção de agentes intencionais artificiais.

ABSTRACT

SANTOS, Guilherme Silveira de Almeida Santos. *Intentionality and Artificial Intelligence*. University Federal in Goiás. Goiânia, 2010.

The concept of intentional stance, a very important concept of Daniel Dennett's philosophical thought, is a central aspect to a naturalized viewpoint of intentionality.

There are three levels to predicting the behavior of an object:

1) The physical stance, at this level, the prediction is based on the physical properties or physical laws.

2) The design stance, at this level, the prediction is based on the function of an object.

3) The intentional stance, at this level, the object is considered a intentional agent, that has belief, thinking and intention.

More over, from Dennett's point of view, the naturalization of intentionality is the way to the possibility of construction of computers that will have intentional behavior. Also, the development of intelligent computers is the objective of Artificial Intelligence (AI) research.

And also, is important to show the refutations of two skeptical arguments that try to prove the impossibility of machines intentionality. Two arguments against some aspects of AI research are the Gödel's theorem argument and the chinese room argument.

The objective of dissertation is show that the concept pf intentional stance is a possibility to construction of artificial intentional agents.

INTRODUÇÃO

A intencionalidade é um problema conceitual que tem ocupado os pensadores desde a tradição fenomenológica até a filosofia da mente contemporânea. O conceito de intencionalidade foi primeiramente abordado pelos filósofos Franz Brentano e Edmund Husserl. Consideremos a tese de Brentano (BRENTANO, 1874/1995): intencionalidade é o que distingue fenômenos mentais dos objetos físicos. Um modo de expressar essa tese clássica consiste em dizer que todo e qualquer fenômeno mental exibe intencionalidade e nenhum fenômeno físico apresenta intencionalidade. Os estados mentais são caracterizados essencialmente por seu aspecto intencional, ou seja, pela propriedade de possuir um certo conteúdo, dirigir-se a, ser sobre ou referir-se a objetos e estados de coisas no mundo. Essa tese, assim como outras propostas desenvolvidas pela filosofia desde então, procura compreender a natureza essencial da intencionalidade dos fenômenos mentais. Como exemplo disso, quando pensamos acerca de árvores ou carros, nossos pensamentos têm um conteúdo, ou se referem às árvores ou carros.

Adicionalmente, o campo de pesquisa científica conhecido como Inteligência Artificial (IA), desde o final da década de 50 tem por objetivo desenvolver sistemas de computador que possam ser considerados inteligentes. Filosoficamente a pesquisa em filosofia da mente implica em saber se podemos, ao menos em princípio, construir um sistema artificial que apresente o que denominamos acima como intencionalidade. A importância filosófica do pensamento de Dennett para a questão da IA consiste em que a posição naturalista defendida por ele considera que computadores inteligentes são possíveis.

Diversas tentativas foram elaboradas pelos teóricos para compreender a intencionalidade e também analisar a questão a respeito da possibilidade de uma intencionalidade artificial. As hipóteses e teorias sugeridas envolvem o que pode ser chamado de naturalização. Isso significa que a filosofia da mente

tenta esclarecer tais problemas tornando-se, parcialmente , um ramo da filosofia da ciência, utilizando os dados e a metodologia das ciências naturais. Como conseqüência da naturalização, os debates filosóficos estão entrelaçados com as abordagens científicas no que diz respeito ao entendimento da intencionalidade e os demais aspectos da mente.

O objetivo da presente dissertação é analisar a concepção naturalizada de intencionalidade apresentada por Daniel Dennett. Pretende-se também analisar determinados aspectos filosóficos da IA para que seja possível compreender as refutações elaboradas por Dennett em relação a dois argumentos céticos em relação à IA, a saber : o argumento do teorema de Gödel e o argumento do quarto chinês.

A filosofia da mente é, em parte, uma área ligada à corrente da filosofia analítica contemporânea. O pensamento de Dennett foi influenciado em grande medida pelo filósofo analítico Gilbert Ryle. Como uma forma de se opor ao dualismo (RYLE,1949), o pensamento de Ryle considerava o problema mente-corpo como um pseudo-problema passível de ser resolvido pela análise da linguagem. Na esteira de Ryle, Dennett considera os estados mentais e a intencionalidade não como entes reais, mas sim como construções teóricas.

Devido ao fato de ser um ponto de vista naturalista, a concepção de Dennett se opõe à tradição fenomenológica. Na teorização de Dennett, é necessário seguir os dados empíricos relevantes para que se possa decidir o que considerar em uma formulação acerca da intencionalidade. A concepção de Dennett deve ser entendida como uma forma de funcionalismo clássico em oposição ao dualismo.

1. Teorias sobre a relação mente-cérebro

Uma possível maneira de apreendermos o pensamento de Dennett no tocante à intencionalidade enquanto forma de estado mental, consiste em fazermos, em um primeiro momento, uma análise de outras abordagens

naturalistas dos estados intencionais. Isso é um passo relevante para a introdução do ponto de vista de Dennett.

No escopo da filosofia da mente, temos primeiramente o que foi denominado de fisicalismo. O fisicalismo se apresentou inicialmente sob a forma do que filosoficamente é chamado de teoria da identidade de tipos (ARMSTRONG,1968). A teoria de identidade compreendia os fenômenos mentais, (bem como as propriedades dos estados mentais, tais como a intencionalidade), como uma simples resposta a uma equação descrita como se segue:

estado mental= estado cerebral (ARMNSTRONG, p 12, 1968)

O significado dessa equação é que para cada estado mental (incluindo dores, pensamentos, intenções, etc.) haveria um estado cerebral que seria identificado pela pesquisa científica. O estado cerebral correspondente ao estado mental consistiria em um padrão de atividade neurológica localizado em algum lugar do sistema nervoso. A teoria de identidade é uma forma não dualista de abordagem dos estados mentais. Um teórico da identidade não atribui a intencionalidade a algo intangível, ou seja, não se deve supor que a intencionalidade consiste em uma ocorrência além do mundo físico, permanecendo além de qualquer abordagem científica.

A dificuldade da teoria da identidade é o seu caráter não parcimonioso: a teoria postula um conjunto demasiadamente complexo de estados cerebrais para todo estado mental. De acordo com a teoria da identidade:

X apresenta intencionalidade se, e somente se, determinados neurônios em seu cérebro estão no estado eletroquímico y.

Ora, segundo Dennett(Dennett, 2006 p 18) há um número imenso de objetos possíveis para se compor um predicado mentalista sobre esses mesmos objetos. Seria inconcebível que, para cada nuvem ou pedra que pudéssemos encontrar, haveria um estado cerebral em uma condição física diferente, estado cerebral que seria diferente para cada um dos objetos do mundo empírico. Não é de forma alguma praticável compor predicados na linguagem da física para identificar todas as nuvens ou pedras existentes no mundo. Se considerarmos, à guisa de exemplo, um determinado despertador digital, poderíamos a princípio indagar o que exatamente o diferencia de quaisquer outros despertadores. Dito de outro modo: qual é a característica física comum entre um despertador analógico e um correspondente despertador digital para que, em razão dessa dada característica física, seja possível dizer " este despertador foi programado para emitir um ruído às oito horas da manhã" ? Ocorre que, o ponto comum entre qualquer despertador é apenas a sua função específica, e não a sua constituição física.

Além da teoria da identidade, foi proposta uma outra forma de fisicalismo para se alcançar uma compreensão da intencionalidade. Essa forma de teorização é chamada de funcionalismo das máquinas de Turing (PUTNAM,1975).Uma máquina de Turing, conforme será explicado no terceiro capítulo, consiste em um procedimento formal de cálculo que fundamenta as operações de qualquer computador possível. Assim como a teoria da identidade, o funcionalismo de máquinas de Turing também nega o dualismo no sentido de não atribuir a intencionalidade a alguma característica não física. A diferença básica entre a teoria da identidade e o funcionalismo de máquinas de Turing consiste em dizer que esse funcionalismo postula que para cada predicado intencional A, existe em correspondência um predicado B expresso em linguagem fisicamente neutra, ou seja, considera-se que o substrato físico, seja ele qual for, não seja fundamental para a teorização.

O que a teoria de máquinas de Turing supõe (PUTNAM,1975). é que os predicados possam ser expressos para especificar funções e relações funcionais. Uma vez que o próprio conceito de máquinas de Turing se refere aos princípios lógicos de qualquer computador, a maneira de se exprimir os predicados são os sistemas de computadores ou programas. A estrutura

funcional de um programa computacional pode ser descrita de forma completamente neutra e abstrata, pois não depende de nenhuma descrição do hardware físico. Em relação à funcionalidade, a linguagem mais neutra possível para descrever as atividades dos computadores é considerar essas mesmas atividades como operações de uma máquina de Turing. Dessa maneira, pode-se descrever matematicamente os estados de um computador digital como os estados e atividades de uma única máquina de universal de Turing, e essa descrição identifica uma determinada máquina de todos os outros computadores ou programas funcionalmente diferentes, mas de forma independente da estrutura física dos computadores. Em síntese, o funcionalismo de máquinas de Turing procura abordar a intencionalidade através da seguinte maneira:

X apresenta a intenção a se e somente se, x realiza uma máquina de Turing TM no estado lógico L.

O significado dessa expressão(PUTNAM,1975). é que para que dois indivíduos quaisquer acreditem ambos na proposição " o céu é azul" eles não precisam ser semelhantes em termos de constituição física, mas somente devem estar em uma condição especificável em linguagem funcional. Os indivíduos devem compartilhar uma descrição de máquinas de Turing de acordo com a qual eles estão ambos em algum estado lógico particular, algo análogo ao fato de dois computadores específicos possuírem o mesmo programa. A tentativa de descrever a intencionalidade e os demais estados mentais de forma física foi substituída por uma forma de redução dos predicados intencionais a meros predicados de máquinas de Turing.

Como corolário dessas considerações, pode-se(Dennett, p 19 2006) fazer uma comparação entre a teoria de identidade e o funcionalismo de máquinas de Turing enquanto formas de teorização fisicalista. Pode-se encarar a teoria de identidade como uma abordagem que postula a semelhança de ocorrências, pios supõe-se que cada evento mental é idêntico a

um evento cerebral individual. Em contrapartida, o funcionalismo de máquinas de Turing é uma forma de abordagem fisicalista particular da intencionalidade no sentido de ser uma teoria que postula a semelhança de tipos abstratos de atividade, ou seja, para cada evento mental há um padrão formal de atividades definidas abstratamente em termos de estados lógicos de um computador. Dessa maneira, a intencionalidade é compreendida como uma atividade formal e abstrata identificável através da linguagem da descrição das máquinas de Turing.

Assim como a teoria da identidade, a abordagem das máquinas de Turing apresenta algumas dificuldades no que diz respeito á compreensão da intencionalidade. Essa abordagem pressupõe que, em princípio, ocorre uma maneira de descrever a intencionalidade (e, por extensão, os estados mentais em geral), como simplesmente estados lógicos de uma máquina de Turing, ou seja, estados funcionais que são pela própria definição neutros em relação ao substrato físico que um dado sistema pode apresentar. De maneira que poderíamos dizer que a semelhança entre duas pessoas que apresentam uma dor consiste no fato de ambas as pessoas estarem no mesmo estado de máquina de Turing. Contudo, embora pode-se dizer que o cérebro é uma forma de computador (Dennett, p 19 2006), não podemos estar seguros que duas pessoas, por apresentarem alguma sensação ou intencionalidade, compartilham, ambas, o mesmo estado lógico de máquina de Turing. Em outras palavras, não ocorre que duas pessoas possam apresentar o mesmo programa de computador, na acepção formal e funcional de um programa de computador, acepção esta que considera irrelevantes considerações acerca de estrutura física das pessoas do exemplo em questão. Essa dificuldade do funcionalismo de máquina de Turing é semelhante ao problema da teoria da identidade, a saber, apresentar um estado mental intencional não implica necessariamente que em uma descrição fisiológica idêntica por parte dos indivíduos que possuem um estado mental.

Uma outra forma de naturalização da intencionalidade é a proposta de John Searle, denominada “naturalismo biológico”. Searle propõe (SEARLE, 1984) que a intencionalidade consiste em um fenômeno irredutível ao vocabulário neurológico. A intencionalidade se deve, nessa concepção, a

propriedades causais do cérebro. Em síntese, a intencionalidade é causada pelo cérebro, mas não é idêntica a o próprio cérebro. Conforme será discutido no terceiro capítulo da presente dissertação, essa hipótese apresenta a dificuldade essencial de não definir o quem são essas propriedades causativas e tampouco como a atividade científica poderia compreendê-las através de um escrutínio experimental adequadamente controlado.

2 O funcionalismo proposto por Dennett

Para tentar suplantar as dificuldades acima mencionadas, a teorização de Dennett procura, antes de tudo, preservar o funcionalismo, asseverando que todo evento mental é um tipo de evento funcional. Ocorre que a teoria de Dennett, lança mão da noção de postura intencional. Essa noção pode ser expressa através do seguinte exemplo:

Um indivíduo A apresenta uma crença intencional z, se for possível fazer a predição de que A tem a crença z. (Dennett, p 20, 2006)

A postura intencional pode parecer à primeira vista uma estratégia tautológica na medida em que de forma redundante parece pressupor o que se pretende explicar, a saber a intencionalidade. Mas a aparente falácia de petição de princípio pode ser esclarecida se considerarmos, por comparação, o caso das máquinas de Turing. Suponhamos que temos duas realizações físicas de uma máquina de Turing . Essas duas realizações específicas podem ser diferentes nos mais diversos aspectos, principalmente em termos de sua incorporação física. Como exemplo dessa afirmação, pode-se dizer que uma das máquinas é realizada por um dispositivo mecânico construído com roldanas, enquanto que a outra máquina em questão é realizada por um processador à base de circuitos eletrônicos integrados. Ocorre que, no momento em que essas duas máquinas estão no mesmo estado lógico, o que elas possuem em comum é somente um sistema de descrição que estabelece que ambas as máquinas são realizações de alguma máquina de Turing específica, sendo que essa descrição permite fazer predições funcionais

acerca dos dois dispositivos. O sistema de descrição apenas diz respeito ao fato dos dispositivos estarem no mesmo estado de máquina. O cerne da questão é que o arcabouço conceitual sobre as máquinas de Turing não está sendo reduzido ou eliminado. Tampouco pode-se dizer que o sistema de descrição acarreta alguma forma de tautologia. O que de nos parece(Dennett, p 20, 2006) é que esse sistema descritivo fundamenta adequadamente todo o discurso sobre as máquinas de Turing, pois aplica a esse discurso regras de estado de máquina, o que por sua vez torna possível atribuir aos dispositivos um papel funcional, bem como também fazer predições sobre esses mesmos dispositivos enquanto formas de realização de máquinas de Turing. A partir do momento em que possamos fazer predições sobre os dispositivos, nossas estratégias de explicações alcançam ima capacidade explanatória adequada. Por analogia, se for possível tornar legítimo o discurso acerca da intencionalidade, não precisaremos procurar reduzir esse discurso a alguma outro discurso que porventura pensemos ser mais coerente. Para a questão filosófica da intencionalidade,, a analogia com as máquinas de Turing é decisiva, conforme sustenta Dennett:

> Supõe-se que os sistemas intencionais desempenham na legitimação dos predicados mentalistas um papel que é paralelo ao papel desempenhado pela noção abstrata de máquina de Turing na atribuição de regras para a interpretação dos artefatos como autômatos computacionais. (DENNETT, 2006, p21,).

Um aspecto interessante a ser salientado sobre a diferença entre o discurso intencional e as máquinas de Turing consiste em que enquanto o conceito de máquinas de Turing é um recurso formal que visa esclarecer noções estritamente exatas pertencentes ao domínio das funções recursivas, bem como também os outros aspectos lógicos- matemáticos da ciência da computação, o discurso intencional (englobando no caso em questão a postura intencional de Dennett) possui em grande medida uma

imprevisibilidade na medida em que o campo que ele procura esclarecer é o domínio das ações e atividades dos agentes intencionais.

Uma maneira de enfatizar a comparação entre a postura intencional e as máquinas de Turing, é através da análise da acima mencionada tese de Brentano. Reiterando o que foi exposto anteriormente, basicamente a tese de Brentano estabelece que o que distingue um estado mental de um mero evento ou objeto físico é a presença da intencionalidade. Todo e qualquer estado mental possível exibe intencionalidade, ou seja, um estado mental apresenta um conteúdo que é sobre ou se refere a objetos ou estado de coisas no mundo. No dizer de Brentano (BRENTANO,1874) “Podemos assim definir os fenômenos psíquicos dizendo que eles são aqueles fenômenos os quais, precisamente por serem intencionais, contém neles próprios um objeto.”

Um modo de expor esse ponto de vista é dizer que os estados mentais sempre são acerca de alguma coisa, desse modo quando um agente exibe intencionalidade, essa intencionalidade é acerca de alguma coisa. A tese de Brentano prescreve, por sua própria definição que todo evento mental apresenta intencionalidade e nenhum evento físico pode apresentar intencionalidade. Conceitualmente, a tese de Brentano é uma forma de teoria que estabelece o aspecto de irredutibilidade da intencionalidade: nenhum evento ou estado mental pode ser reduzido ao domínio físico pois todo estado mental apresenta intencionalidade. Todavia, consideremos uma asserção da tese de Brentano, qual seja, a afirmação de que todo estado mental exibe intencionalidade.. Essa asserção é parecida, no dizer de Dennett (Dennett, p 21 2006) com a tese de Church, proposta pelo matemático Alonzo Church (HOFSTADTER,1979). Essa tese fundamenta a computação estabelecendo que os procedimentos matemáticos recursivos podem ser executados por uma máquina de Turing. Desse modo, qualquer conjunto finito de etapas simples ou qualquer procedimento definido como um algoritmo capaz de realizar uma função matemática, pode ser implementado por um computador denominado máquina universal de Turing. Na medida em que depende de uma noção não –formalizável de um procedimento eficiente, a tese de Church não pode ser demonstrada como verdadeira, embora seja aceita de modo geral pela comunidade de pesquisadores envolvidos na pesquisa de ciência da

computação. A aceitação da tese de Church deriva do fato de ser possível, através dessa tese, alcançar uma definição reducionista de uma noção matemática, definição essa que, ao ser relacionada ao funcionamento de um computador teórico, pode proporcionar grande capacidade explicativa. Por comparação, a assertiva de que todo estado mental apresenta intencionalidade (segundo essa concepção, a intencionalidade é a característica básica dos sistemas intencionais), poderia oferecer uma redução dos termos mentalistas. A veracidade da afirmação de que a intencionalidade é a marca do mental poderia, teoricamente, permitir que os estados mentais sejam abordados não através de noções intuitivas e conhecimento tácito das experiências de senso comum, oferecendo aos pesquisadores um escopo de conceitos sistematizados e melhor fundamentados.

3 Estrutura da dissertação

No primeiro capítulo da dissertação pretende-se analisar a postura intencional, comparando-a com as demais estratégias de abordagem do comportamento das entidades, tais como as posturas física, e de projeto. A postura intencional é uma abordagem funcionalista na medida em que descreve a intencionalidade através de uma linguagem preditiva que é neutra em relação ao fisicalismo tradicional. A estratégia de postura intencional pressupõe que os componentes físicos estruturais de uma entidade não são fundamentais para tratarmos essa entidade como um sistema intencional. A noção de neutralidade em relação ao fisicalismo se justifica na medida em que um objeto físico pode ser tratado como um sistema intencional mesmo que nem todos os sistemas intencionais possam ser efetivamente realizados, e mesmo que nem todos os predicados mentalistas possam ser identificados com aspectos de um sistema intencional irrealizável..

A estratégia da postura intencional, embora seja uma formulação teórica fisicalista, não pretende legitimar os predicados intencionais através de uma identificação dos termos mentalistas com características e entidades reais. A postura intencional de Dennett considera que todo estado mental é também um evento funcional. Contudo, os predicados mentais são somente instrumentos para se realizar predições do comportamento das entidades. A noção de

postura intencional não considera os estados mentais como reais no mesmo sentido dos elétrons e neurônios. No que diz respeito aos estados mentais bem como à intencionalidade, a postura intencional não é uma forma de identificação ontológica desses estados com algum objeto físico discernível.

O segundo capítulo da presente dissertação pretende elaborar uma reflexão filosófica a respeito da IA. Pretende-se em síntese mostrar como a hipótese de sistemas artificiais apresentarem intencionalidade é coerente, constituindo-se em uma questão de enorme interesse epistemológico. A abordagem funcionalista da intencionalidade, adotada por Dennett, consiste no ponto central para mostrarmos que um computador pode apresentar comportamento intencional: tudo o que é preciso para a intencionalidade de qualquer sistema é somente a instanciação de um programa adequado, independentemente dos componentes materiais do sistema.

A pesquisa de IA recebeu críticas por pesquisadores que não consideram possível que um computador possa de fato apresentar intencionalidade. No terceiro capítulo da dissertação pretende-se elaborar uma refutação dos já mencionados argumentos céticos em relação à IA, o argumento do quarto chinês e o argumento do teorema de Gödel. Cada um desses argumentos serão expostos e analisados com o objetivo de mostrar o modo através do qual ambos os argumentos incorrem em erros filosóficos. Pode-se dizer que o argumento do quarto chinês comete uma falácia conceitual, em contrapartida o argumento do teorema de Gödel um erro lógico-matemático. Será mostrado que esses argumentos céticos em relação à pesquisa de IA não demonstram que um sistema artificial de computador não possa apresentar uma forma de intencionalidade, pois o que torna a intencionalidade possível é a realização funcional de um programa. Caso uma entidade física realize uma organização funcional apropriada, pode-se considerar que essa entidade possa de fato ser considerada um agente intencional.

Por fim, na última parte da dissertação serão feitas as considerações finais. Pretende-se argumentar acerca da importância do pensamento de Dennett no tocante à compreensão da intencionalidade e também enfatizar o

fato de que tal pensamento proporciona um esclarecimento sobre a pesquisa de inteligência artificial.

CAPÍTULO 1: A POSTURA INTENCIONAL

O conceito de sistema intencional utilizado por Dennett significa o êxito prognosticado ao se aplicar a um determinado sistema em questão uma postura intencional. O uso da postura intencional se justifica quando o sistema em questão é demasiado complexo o bastante a ponto de atribuir-lhe termos intencionais, tais como crenças, desejos, medos , objetivos, etc.. Basicamente, um sistema pode ser chamado de intencional se pudermos aplicar os acima mencionados estados mentais na explicação e predição de seu comportamento.

Um ponto importante a ser enfatizado aqui consiste em que a intencionalidade não é algo que não possa ser reduzido ao domínio físico; De acordo com a tese de Brentano, a intencionalidade é o que distingue o mental do físico, significando que o intencional é considerado irredutível ao físico. O que distingue o mental do físico é o fato de que um estado mental, por sua própria definição, apresenta um conteúdo que se dirige, é sobre, ou se refere a objetos e estados de coisas no mundo. Conforme o que foi anteriormente mencionado, essa propriedade de um estado mental apresentar um conteúdo caracteriza a intencionalidade.. Apesar disso, Dennett (Dennett, 2006 p 34) não entende a intencionalidade como uma propriedade intrínseca dos estados mentais. Algo é um sistema intencional, de acordo com Dennett, apenas em relação a uma postura adotada em relação ao sistema em questão. Desse modo, demonstraremos que Dennett acredita que o conceito de intencionalidade pode ser compreendido ou reduzido ao escopo das ciências físicas, de maneira diametralmente oposta a abordagem de Brentano e a fenomenologia tradicional, que trata a intencionalidade como uma capacidade mental irredutível.

1. A postura intencional e as abordagens física, e de projeto

Consideremos, à guisa de exemplo, um sistema de computador capaz de jogar uma partida de xadrez. De acordo com Dennett(Dennett, 2006 p 35), há três modos distintos de considerar esse sistema ao se tentar entender seu comportamento. Primeiramente, há o que podemos chamar de *postura física*. Esta abordagem usualmente é adotada quando estamos diante de um sistema tão simples que apenas o conhecimento das regularidades físicas envolvidas é suficiente para se compreender o comportamento do sistema, tal como a trajetória de uma pedra caindo em direção a superfície terrestre. Basta, em tal situação, que se considere a ação da força gravitacional para se obter uma predição eficiente. Teoricamente, a postura física pode ser aplicada ao referido sistema de computador, levando-se em conta o conhecimento de propriedades de materiais semicondutores e correntes elétricas. Ocorre que o sistema de computador possui sub-rotinas demasiado complexas para ser compreendido de forma puramente física.

Além da postura física, há o que Dennett(Dennett, 2006 p 36) denomina de *postura de projeto*. Essa abordagem pode ser utilizada quando as partes projetadas do computador são conhecidas. È possível predizer cada movimento que o computador fizer durante a partida de xadrez apenas seguindo as instruções de seu programa. As predições a partir da postura de projeto alcançarão êxito ao assumirmos que cada parte funcione de modo adequado. As sub-rotinas elaboradas a partir de cada diagrama dos resistores e capacitores possuem cada uma delas uma determinada função específica. Pode-se predizer efetivamente o comportamento do computador se supormos que cada sub-rotina ou circuito cumprirá exatamente a função que o engenheiro designou. As partes a serem consideradas na postura de projeto podem variar desde sub-rotinas de estratégias de jogo ou mesmo analisadores de operações aritméticas.

Geralmente, usa-se a postura de projeto no tratamento de entidades como um despertador, ao assumirmos que ele consiste em um aparelho cujo

funcionamento foi deliberadamente projetado para emitir um sinal em um determinado momento especificado no tempo pelo usuário. Note-se que, não é necessário considerar as propriedades físicas dos materiais que constituem o aparelho, mas somente é preciso fazer predições referentes aos aspectos funcionais do sistema. Essa concepção funcionalista considera que o que é relevante para a compreensão de um objeto complexo, como por exemplo o despertador do exemplo em questão, são apenas as considerações sobre o que o objeto pode fazer, e não considerações acerca dos componentes estruturais envolvidos na construção do objeto.

Por fim, ao tratarmos com o comportamento de uma entidade complexa temos um recurso chamado por Dennett de *postura intencional (intentional stance)*. Através dessa estratégia, tratamos o objeto como algo capaz de comportamento racional e postulamos que ele possua algumas crenças, dados os seus objetivos(DENNETT,1987). Essa postura se justifica quando um sistema é muito complicado para ser abordado pelas posturas física ou de projeto. A única forma praticamente aplicável a um sistema computadorizado projetado para efetuar partidas de xadrez é predizer suas respostas supondo que o sistema é um agente intencional capaz de realizar a opção mais racional conforme as normas do jogo de xadrez. Deve-se esperar que esse sistema:

1. funciona de modo como foi projetado para fazer determinada tarefa;
2. consegue escolher, entre vários caminhos de ação, a jogada mais adequada na partida de xadrez.

Essas pressuposições consistem em um modo confiável, embora passível de erros de avaliação, quando não há outra maneira de se predizer o comportamento da entidade em questão. A postura intencional é relativa a um escopo de informações e postulados sobre objetivos, como também é vulnerável a falta de critérios a respeito do que seria em dado momento uma ótima estratégia, sendo que essa vulnerabilidade não pode ser totalmente

prevista a partir da perspectiva intencional. Considerando o sistema como projetado com o objetivo de vencer a partida e levando-se em conta de que ele possui informações em sua programação, deve-se questionar qual a melhor estratégia de jogo considerando certas informações (relativas a posição inicial das peças no tabuleiro de xadrez, etc) em relação a um certo estado de coisas específico diante dos quais o sistema deve procurar alcançar êxito, empreendendo medidas que procurem reduzir as diferenças entre seus objetivos e as variadas situações do jogo de xadrez.

A pressuposição da racionalidade é o que caracteriza a postura intencional. As predições inferidas podem ser chamadas de intencionais na medida em que tratamos a entidade como um sistema intencional, assumindo que ele disponha de certas informações e que também seja provido de determinados objetivos. Assim, a noção de racionalidade se justifica na medida em que em que essas suposições sejam adequadas para se prever o comportamento do sistema. Pode-se chamar a informação que o sistema possui de *crença*, assim como seus objetivos de *desejos*. A intencionalidade consiste no êxito alcançado no prognóstico do comportamento do sistema a partir da postura intencional adotada. O significado da atribuição de racionalidade consiste na correspondência entre nossas predições acerca das ações do sistema e o subseqüente comportamento observado do sistema em causa.

Pode-se alegar que a noção de postura intencional conforme especificada por Dennett(Dennett, 2006 p 38) levanta dúvidas no que concerne ao fato de se saber se o sistema em questão realmente apresenta alguma forma de intencionalidade. Tais objeções seriam a princípio injustificadas, pois, segundo Dennett a abordagem da postura intencional não diz que um sistema *realmente* possua intenções, mas tão somente que se possa prever aproximadamente seu comportamento atribuindo crenças e intenções a um determinado sistema.

> [...] se chamamos aquilo que atribuímos ao comportamento de crenças, ou de análogo de crenças, ou de complexos de informações, ou do que seja de intencional, isso não faz nenhuma diferença para a natureza do cálculo feito com base em tais atribuições. Vamos chegar às mesmas predições quer pensemos em termos das crenças e desejos do computador, quer pensemos em

> termos de armazenamento de informação ou especificação de objetivos. (Dennett, 2006 p 38) .

Recorrendo à analogia dos computadores que jogam xadrez, freqüentemente seu projeto é de tal maneira complexo que só através da postura intencional pode-se de fato tentar compreender as estratégias desenvolvidas ao longo do jogo. A postura intencional é um tratamento puramente funcional e pragmático, não fazendo em nenhum momento considerações acerca da base física de que é feita a estrutura dos componentes do sistema, ou seja, tem-se uma acepção puramente formal e abstrata da atribuição dos estados mentais a um sistema: somente o *software* é fundamental e não o *hardware*. É possível mesmo alternar a adoção de posturas conforme mudamos de estratégia de predição de comportamento. Desse modo, para fins de manutenção estrutural, utilizamos a postura física em relação a um computador. Em contrapartida, se vamos jogar uma partida de xadrez contra o computador, a única postura adequada é a postura intencional. Adotar a postura intencional não equivale, segundo Dennett(Dennett, 2006 p 40), em tratar uma entidade de maneira equivalente a um agente consciente. Trata-se, antes de tudo, no fato de que alguns sistemas são complexos o bastante para que a predição explicativa de seu comportamento possa ser feita mais facilmente ao se considerar o sistema em questão como um agente racional imbuído de crenças e desejos.

Ao elaborar a noção de postura intencional, Dennett(Dennett, 2006 p 41) relaciona essa formulação com outras teorias comportamentais aplicáveis aos mais diversos sistemas, sejam os mesmos humanos ou computadores. A princípio, pode-se notar que as assim chamadas explicações de senso comum, ou seja, as estratégias que usualmente utilizamos em situações cotidianas, são intencionais na medida em que pressupõem a racionalidade dos sistemas considerados. Não se espera que um indivíduo reaja irracionalmente a alguma situação, contudo, caso isso ocorra, procuramos sempre ajustar nossas pressuposições iniciais. Nos eventos da vida cotidiana, estamos acostumados a pressupor a racionalidade dos indivíduos com os quais interagimos. Dessa

maneira, se as ações de um dado indivíduo não se ajustam com as nossas predições, ajustamos as condições dessas predições (por exemplo, pode-se considerar que o indivíduo em questão fale um idioma que nós não compreendemos) assumindo ainda sim que o indivíduo é capaz de agir racionalmente. Se, ao acumularmos uma quantidade razoável de conhecimentos referentes aos padrões de resposta de uma entidade, ainda assim constatarmos que a mesma permanece inescrutável de acordo com a postura de intencional (não estamos conseguindo fazer predições satisfatórias do comportamento dessa entidade através da postura intencional), concluímos que a postura de projeto ou mesmo a postura física podem ser suficientes para uma predição comportamental satisfatória.

Nossas explicações são também intencionais quando tentamos prever o comportamento dos animais . . (DENNETT, 2006, p 41)Se observarmos um rato em um ambiente onde ele claramente pode notar a presença de um gato, podemos assumir a postura intencional em relação ao rato e prever que ele evitará se aproximar do gato. Fazemos isso supondo que o rato possui algo que poderíamos chamar de análogos de crenças e objetivos e, desse modo, concluímos, em relação a situação acima apresentada, que o rato tem como objetivo evitar o perigo. Qualquer que seja a teorização psicológica formulada para explicar em termos científicos o comportamento do rato, espera-se que ele se comporte da maneira descrita acima.

Há uma questão conceitual que pode ser feita nesse ponto. Podemos perguntar se os animais, ao serem caracterizados como sistemas intencionais, são capazes de, se não compreender, ao menos seguir irreflexivamente os fatos lógicos. A definição de alguma coisa como um sistema intencional envolve a seguinte significação:

> A pressuposição de algo como um sistema intencional é a pressuposição que ele é racional; isto é, não se chega a nada com a pressuposição de que a entidade x possui as crenças p, q, r,..., a menos que se suponha também que x acredita naquilo que se segue de p, q, r,...; de outro modo, não há como descartar a predição de que x, tendo em conta suas crenças p, q, r,... vá fazer algo completamente estúpido e, se não pudermos eliminar *essa* predição,

> não vamos ter adquirido nenhum poder preditivo que seja. .
> (Dennett, 2006 p 42)

Pode-se esperar, à guisa de exemplo, que o rato mencionado anteriormente siga algumas conclusões lógicas no sentido de que atribuímos a ele as seguintes crenças:

!) Há um gato;

2) Se há um gato, devo evitar me aproximar dele.

Nossas predições são otimizadas na medida em que constatamos a capacidade do rato de seguir essa conclusão. Assim, no que concerne à postura intencional aplicada ao rato, podemos lhe atribuir algumas proposições e regras de inferência lógicas, como, por exemplo, as do tipo “se p, então q”.

Em um sistema intencional perfeitamente racional todas as verdades lógicas apareceriam de forma ideal entre todo o conjunto de suas crenças. Contudo, um sistema efetivamente existente no mundo é imperfeito e apenas aproximadamente racional. Não se deve atribuir a um sistema intencional real todas as verdades lógicas como constituintes do seu escopo de crenças. Cumpre-se notar que nem todas as regras de inferência de um sistema intencional real serão válidas; nem todas as suas crenças que possibilitam inferência serão verdades lógicas. Através da experimentação, assumindo um método verificacionista, os erros de um dado sistema intencional podem ser detectados. Se estivermos diante de um ente cuja fidelidade a uma regra de inferência variasse em diferentes circunstâncias, excluiríamos essa regra em relação ao ente. As explicações intencionais que tentamos elaborar se tornariam cada vez mais difíceis, pois não observaríamos as ações comportamentais adequadas segundo as crenças e objetivos que supomos que a entidade deveria possuir, dadas as regras lógicas que pressupomos que estejam sendo compreendidas. O resultado desse impasse é que não mais adotamos a postura intencional, e sim as posturas física ou de projeto.

A mudança de postura para se atingir explicações mais confiáveis e melhor fundamentadas do comportamento de sistemas imperfeitamente racionais é,

segundo Dennett(Dennett, 2006 p 44), uma forma de alcançar uma teorização que explique a intencionalidade. Quando um pesquisador atribui a algum componente estrutural de um sistema definições como por exemplo sinais, comandos ou mesmo intencionalidade, ele postula com os comandos ou sinais algo que pode ser chamado de decodificador de sinal; caso contrário a noção de sinal não terá sentido. Todo esse arcabouço teórico do pesquisador deve ser compensado através da descoberta dos fatos básicos que possibilitam a intencionalidade. Caso contrário, os sinais e decodificadores não são eliminados e ocorre uma petição de princípio: tem-se uma explicação incompleta que tautologicamente *pressupõe* a intencionalidade na teoria científica, sem que se preocupe em explicar em que a intencionalidade consiste. Esse erro filosófico é denominado falácia do homúnculo. No dizer de Dennett (Dennett, 2006 p 44) " onde quer que uma teoria se baseie em uma formulação que contenha as marcas lógicas da intencionalidade, lá um pequeno ser humano está escondido".

Encontrar uma abordagem satisfatória da intencionalidade constitui uma forma de assegurar uma compreensão adequada do comportamento das entidades complexas . Se, entre os diversos conceitos de uma teoria, houver algum aspecto no qual a intencionalidade esteja sendo assumida por princípio, a explicação formulada não é satisfatória.

2. Postura intencional e o behaviorismo

Essas considerações podem ser utilizadas para se demonstrar as vantagens da postura intencional em relação às explicações behavioristas do comportamento. Os pensadores behavioristas (SKINNER, 1979) consideravam incorreta a utilização de termos intencionais em suas explicações do comportamento. O behaviorismo, enquanto sistema de pensamento na psicologia, considerava as explicações intencionais como empiricamente desprovidas de significado, sendo que apenas o comportamento externo

observável era relevante para a formulação das bases de uma teoria psicológica científica. Deve-se notar que explicações tendo por base a postura intencional são significativas no sentido em que fazem hipóteses que, ao menos em princípio podemos testar. Se perguntarmos a algum conjunto de pessoas razoavelmente instruídas qual a solução da seguinte operação matemática:

25 + 10

Podemos estar certamente confiantes que a maioria das pessoas dirá como resposta o número 35. Caso alguém porventura levante dúvidas ou expresse algum ceticismo a respeito dessa predição, podemos testá-la facilmente na prática. O conteúdo empírico que tal predição aparenta possuir decorre justamente do fato de que se trata de uma predição verificável. Embora a questão científica sobre uma teoria psicológica do comportamento não esteja sendo completamente resolvida por uma predição intencional, podemos notar que essa predição funciona. A razão pela qual uma predição intencional funciona é que qualquer sistema intencional possível é normalmente estruturado para dar respostas corretas em situações apropriadas. Uma predição intencional, prevendo o comportamento através da atribuição de crenças e desejos, constitui um componente fundamental de qualquer teoria psicológica que uma ciência empírica deve formular não apenas com a finalidade de compreender a cognição humana, mas também para a devida elucidação dos sistemas intencionais em geral.

Para determinarmos que um objeto qualquer é feito de carbono ou silício, é suficiente apenas levar em conta a evidência empírica disponível acerca da estrutura físico-química do objeto em causa. Ocorre que, no caso da postura intencional(Dennett, 2006 p 45), decidir com base na evidência física ou os dados empíricos sobre se algo é ou não um sistema intencional, é irrelevante, uma vez que a predição baseada na postura intencional não depende da constituição física de um objeto. O substrato físico de um objeto não é a base da aplicação da postura intencional, e sim os seus aspectos funcionais.

A abordagem behaviorista tentou eliminar a linguagem intencional, substituindo-a por uma metodologia que considerava que um sistema específico pura e simplesmente reagia a dados de entrada ou estímulos, gerando padrões de saída ou respostas, sem considerações a noção de estados mentais. Os teóricos behavioristas consideravam que os termos mentalistas eram cientificamente incoerentes, como o dualismo cartesiano:

> [...] ao transferir o comportamento humano para um mundo de dimensões não físicas, os pesquisadores mentalistas ou cognitivistas formularam as questões fundamentais de modo absolutamente insolúvel. (SKINNER, 1974, p.121)

O fato é que tal metodologia não alcançou predições tão satisfatórias quanto as predições feitas pela postura intencional. Não se pode afirmar que o comportamento de um sistema tenha sido elucidado por uma explicação baseada apenas em estímulos/respostas.

Consideremos as assim chamadas “caixas de Skinner”, elaboradas e utilizadas pelo pesquisador behaviorista B.F.Skinner. Essas caixas são dispositivos de estudo e controle experimental do comportamento em que um dado organismo – um dos exemplos clássicos da literatura behaviorista é um pombo – é colocado em um recipiente onde uma alavanca é pressionada como resposta comportamental a um estímulo físico previamente especificado pelo pesquisador (SKINNER, 1979) O behaviorista prediz que, após um treinamento sistemático, uma resposta comportamental ou reflexo condicionado é observado de forma tal que pode ser previsto e quantificado segundo a análise científica usual. Pode-se dizer que essa reposta tem uma especificidade definida não intencionalmente, devido não a alguma característica do organismo, mas antes devido a um aspecto da função que a caixa realiza. Para ilustrar esse ponto consideremos novamente o exemplo da operação matemática citada acima. Suponhamos que uma pessoa seja introduzida na caixa onde em seu campo visual está um cartão com a operação 25 + 10 e também dois cartões adjacentes com as possíveis respostas 35 ou 36. Se a pessoa for questionada sobre qual dos cartões de resposta é a solução exata da operação, é previsivelmente correto que o

cartão 35 será marcado como resposta, independentemente de um condicionamento antecipadamente aplicado. Essa predição não evita o vocabulário intencional:

> Trata-se, de uma *predição intencional* colocada implicitamente devido as restrições do dispositivo, de modo que a pessoa na caixa somente pode realizar a única ação intencional coerente com a questão matemática proposta. (DENNETT, 2006 P 46)

Em suma, as caixas de Skinner não substituem o intencional por algum conjunto de leis gerais do comportamento; trata-se de que o vocabulário intencional permanece disfarçado nas experiências behavioristas, pois as caixas de Skinner são estruturadas de forma que só é possível realizar uma ação adequada, gerando a ilusão de que essa ação não é intencional. O ponto central dessa digressão realizada acima é que a metodologia behaviorista, bem como também a estratégia de postura intencional, considera o comportamento externo das entidades como um fator decisivo para a confirmação das predições. Ocorre que, enquanto o behaviorismo fundamenta seus métodos em meras respostas a estímulos (como por exemplo, um cachorro que saliva em resposta ao estímulo do som de um sino), a postura intencional fundamenta-se na noção de atribuição de crenças e objetivos a um sistema.

3. Linguagem e postura intencional

A noção de postura intencional também apresenta vantagens como um recurso filosófico muito eficiente para a análise e compreensão dos sistemas intencionais dotados de linguagem.. Isso se deve ao caráter formal da estratégia de postura intencional(Dennett, 2006 p 49). Essa estratégia abstrai dos detalhes circunstanciais acerca da composição ou constituição de uma entidade qualquer, considerando que, independente de sua natureza ou forma, uma entidade pode ser considerada um sistema intencional.

Uma capacidade interessante de alguns sistemas intencionais é a utilização da comunicação simbólica ou linguagem. Dentre o conjunto dos sistemas intencionais(Dennett, 2006 p 49), os sistemas dotados de linguagem são relevantes para avaliarmos algumas expressões intencionais ou termos mentalistas, como as crenças. Um sistema que pode se comunicar apresenta considerável importância no que diz respeito a atribuição a esse sistema de termos como crenças e desejos, pois se não houvesse a linguagem a pressuposição de racionalidade não seria possível e a postura intencional deixaria de ter seu poder explanatório.

Uma vez que a linguagem é uma capacidade que se desenvolveu na natureza pelo processo evolutivo, ela deve ser encarada como uma habilidade sujeita às restrições do entorno ambiental. Assim sendo, é contraproducente atribuir crenças inapropriadas ao ambiente de um agente. As crenças atribuídas devem atender a dois aspectos(Dennett, 2006 p 51):

1) devem se compatíveis como o entorno físico específico a que o sistema intencional está inserido;
2) o sistema intencional deve responder adequadamente às crenças;

Se a capacidade de apresentar crenças corretas e agir adequadamente em relação a essas crenças, através das inferências corretas, for um aspecto relacionado à sobrevivência, então pode-se concluir que houve na natureza uma seleção por sistemas intencionais mais racionais. Um sistema intencional que não faz as inferências corretas não poderia sobreviver:

> Há um encorajamento em Darwin. Se o espaçamento inato de qualidade é um traço ligado geneticamente, então o espaçamento que fez as induções mais bem sucedidas teve a tendência de predominar através da seleção natural. As criaturas equivocadas em suas induções têm uma patética, mas louvável, tendência de morrer antes de reproduzir sua espécie. (QUINE,1977, p. 126)

Determinada crença de um agente seria favorecida evolutivamente se fosse, em média, uma crença verdadeira . Ainda que possa ter havido criaturas possuidoras de sistemas ineficazes e inverídicos, esses sistemas não seriam favorecidos pela seleção natural, devido à sua própria falta de correspondência com os fatos da realidade, ou seja, devido à irracionalidade desses sistemas. De forma lógica ou conceitual, um sistema de crenças falsas é uma incoerência. Essas considerações acerca da coerência com a realidade se aplicam no tocante à evolução da comunicação simbólica. Um sistema de comunicação falso seria uma inutilidade evolutiva. Somente um sistema de comunicação baseado em crenças corretas seria favorecido pela seleção natural:

> A faculdade de comunicação não se estabeleceria na evolução a menos que fosse, em grande medida, uma faculdade de transmitir crenças verdadeiras, o que significa apenas: a faculdade de alterar outros membros da espécie na direção de uma melhor constituição. . (Dennett, 2006 p 51)

Para que possa haver, em princípio, agentes caracterizados como sistemas racionais que utilizam entre si um sistema lingüístico de comunicação simbólica é necessário que determinadas condições lógicas possam ser satisfeitas, a saber(Dennett, 2006 p 51):

1) Frequentemente, se x acredita em p, então p é verdadeiro;
2) Frequentemente, se x reconhece que p, então x acredita em p;

Essas condições lógicas são condições para que um sistema de crenças encontre aplicação no conjunto de crenças de um agente racional. Se existem evidências em prol das crenças, pode-se dizer que ocorre uma coerência com os fatos. Mas, além disso, é necessário que exista indícios de que o agente em causa acredite de fato em suas crenças, o que consiste em uma garantia para avaliarmos seus enunciados. Desse modo, através da avaliação de provas e averiguação do comportamento externo de um agente, pode-se verificar se suas declarações são verdadeiras e também se o agente acredita de fato em suas próprias declarações.

Outro aspecto relacionado às crenças e enunciados lingüísticos(Dennett, 2006 p 52) é a interligação entre suas crenças e objetivos. Essa interligação impossibilita definições não intencionais dos termos mentais. Considere o caso de uma pessoa que esteja bebendo um copo d'água. Essa ação é um indicativo de que a pessoa está com sede, mas somente se pressupormos que a pessoa deseja saciar sua sede. Se questionarmos a pessoa sobre se ela deseja saciar sua sede, o fato de ela beber um copo d'água pode ser uma evidência, porém somente se admitirmos que a pessoa acredita que pode saciar sua sede consumindo água. Caso perguntemos a pessoa se ela acredita que consumir água vai saciar sua sede, sua resposta dependerá da honestidade de seus enunciados.O problema consiste em saber quais as inferências, entre todas as inferências possíveis, são adequadas para elucidar o comportamento do agente. Essas dificuldades podem ser resolvidas se considerarmos que, normalmente, os enunciados de uma pessoa são indicativos de suas crenças e ações devido ao fato de que, em casos gerais, as crenças e desejos que um agente possui são as crenças corretas que o agente deve possuir nas circunstâncias apropriadas. A garantia dessa solução se baseia no processo de evolução que favoreceu a formação de sistemas racionais.

No tocante aos sistemas intencionais dotados de linguagem, podem ocorrer problemas em relação à avaliação de suas crenças(Dennett, 2006 p 52). Se um agente reconhecer crenças que são em grande medida desmentidas pelos indícios disponíveis, ou se essas crenças se contradizem ou se não estão em acordo com outros reconhecimentos feitos pelo agente, então temos uma falha em relação à atribuição de racionalidade a esse agente, sendo que essa falha pode tornar a postura intencional inapropriada com forma de prevermos o comportamento do agente.

Ainda que nenhum sistema intencional seja ideal, no sentido de possuir *todas* as crenças verdadeiras possíveis e também acreditar em cada conseqüência logicamente coerente de suas crenças, para que um sistema intencional se qualifique como racional ele deve conter crenças verdadeiras e também efetuar ações adequadas ao seu sistema de crenças. O fundamento para se verificar a

atribuição de crenças a um agente não se baseia, contudo, em dados introspectivos fenomenológicos, pois esses dados são essencialmente privativos, baseados na perspectiva de primeira pessoa. É preciso que os dados relevantes sejam comunicados, e somente através de um processo lingüístico pode-se transmitir tais dados, permitindo seu exame objetivo, de acordo com a perspectiva de terceira pessoa. Desse modo, a linguagem é um recurso de suma importância para a avaliação das expressões intencionais. Ainda que um enunciado possa estar sujeito à enganos, pode-se estabelecer como certo que, devido à seleção natural, sistemas intencionais racionais, capazes de comunicar crenças verdadeiras através da linguagem, se desenvolveram no mundo real.

CAPÍTULO 2: REFLEXÕES ACERCA DA INTELIGÊNCIA ARTIFICIAL

No capítulo anterior foi apresentada a postura intencional de Dennett como uma proposta naturalizada de intencionalidade. Neste segundo capítulo, pretende-se mostrar as conseqüências do pensamento de Dennett para a corroboração filosófica da hipótese de dispositivos artificialmente construídos serem capazes de apresentarem comportamento intencional. Os pesquisadores de Inteligência Artificial (IA) conseguiram desenvolver sistemas que apresentam desempenho considerável em tarefas complexas, de modo que é razoável considerar que a elaboração e execução de um intencionalidade artificial é compatível com uma visão científica da realidade.

1. Funcionalismo e inteligência artificial

Partindo de uma premissa materialista e naturalista, Dennett considera que a melhor razão para se pensar que as máquinas podem, em princípio, apresentarem pensamento e intencionalidade consiste no fato de que a postura intencional basea-se em uma abordagem naturalista e mecanicista da intencionalidade.

Conforme o exposto acima, máquinas podem, em tese, possuir intencionalidade. Essa possibilidade não é uma violação de alguma lei física, tal qual a construção de um moto-contínuo, dispositivo esse que é considerado impossível pelos físicos pois não está de acordo com os princípios da termodinâmica.

Para Dennett, a intencionalidade é uma noção funcional, ou seja, para adotarmos a postura intencional em relação a qualquer entidade é irrelevante considerarmos a constituição física de que a entidade é constituída. Essa

perspectiva funcionalista é uma justificativa para a possibilidade da intencionalidade em máquinas.

Como uma teoria da filosofia da mente, o funcionalismo prediz que a intencionalidade não é definida em termos do substrato físico-qímico , mas sim através de relações causais (PUTNAN, 1975). Uma outra forma de definir o funcionalismo é dizer que o que faz com que algo tenha intencionalidade ou mente não é o material de que esse algo é feito, mas sim o que tal coisa pode fazer. Se considerarmos um sistema de IA devidamente projetado e verificarmos que ele apresenta um desempenho tão considerável quanto o de uma inteligência natural, poderemos seguramente concluir que o sistema de IA é uma forma de inteligência tão verdadeira quanto qualquer outra possível.

A definição da inteligência em termos de relações causais levantou críticas ao funcionalismo e, por extensão, ao programa de pesquisa em IA. O argumento pode ser expresso da seguinte maneira, a saber:

Um sistema artificial é, por definição, inorgânico, e a inteligência só é possível a partir de sistemas orgânicos.(DENNETT, 2006, p 271)

Essa alegação está baseada no pressuposto de que os materiais orgânicos são capazes de realizar funcionalmente estruturas com capacidades não reproduzíveis em componente inorgânicos. Supõe-se que mesmo que uma duplicata inorgânica isomórfica a um ser humano seja construída, seus estados mentais intencionais não seriam genuínos, pois só uma duplicata bioquímica, construída com compostos orgânicos, seria capaz de apresentar expressões intencionais. Essas afirmações são, contudo, uma forma de vitalismo (DENNETT,1996). O funcionalismo, por sua vez, é uma oposição ao vitalismo no sentido de não considerar que supostas propriedades intrínsecas dos materiais são relevantes para a inteligência:

> A única razão pela qual o material utilizado na construção de um sistema funcional é importante consiste em que, devido a motivos relacionados a fatos bio-históricos, os materiais devem ser compatíveis com os corpos preexistentes no sistema. (DENNETT, 1996 p 75)

O significado disso é que o material utilizado na elaboração do sistema é relevante não porque contenha em si mesmo propriedades causativas cruciais para o mental, mas sim porque os componentes de um sistema são capazes de *suportar* uma organização funcional.

A intencionalidade deve ser encarada, na acepção funcionalista, como algo formal e abstrato, cuja existência não é definida em termos de uma estrutura física específica (DENNETT;HOFSTADTER, 1981). De acordo com o funcionalismo, se um sistema de IA instanciar um programa adequado, ele apresentará intencionalidade. Caso esse sistema venha a ser efetivamente construído, poderemos verificar seu comportamento intencional através da interação com o sistema.

2. Postura intencional e máquinas

Os sistemas projetados pela pesquisa em IA são sofisticados o bastante para só podermos compreender suas ações através da postura intencional. Mesmo que tenhamos um entendimento completo de um computador em termos de circuitos e outros detalhes físicos, não podemos, por motivos práticos, abrir mão de atribuir aos sistemas computadorizados uma certa racionalidade, através de expressões intencionais.

Um problema de projeto em IA começa com uma questão intencionalmente colocada. Os pesquisadores tentam resolver o problema de fazer o sistema executar uma tarefa, como por exemplo o jogo de xadrez, e então definem a questão em termos de vocabulário intencional, considerando que o computador sabe ou decide sobre a melhor linha de ação possível. Mesmo a posse e transmissão de informação de um computador é descrita em termos intencionais.

Uma vez definido o problema, os pesquisadores concedem ao sistema de computador uma capacidade racional, admitindo que se o sistema conseguir resolver o problema, pode-se dizer que em certa medida ele pôde “entender” a situação. A primeira etapa dos pesquisadores é dividir o sistema a ser projetado em partes menores, sendo que cada um desses pequenos subsistemas do computador recebe cada um deles uma tarefa definida em termos de expressões intencionais. Cada uma dessas subdivisões é um conjunto heterogêneo de avaliadores, discriminadores de padrões, verificadores, e assim por diante. Esses subsistemas podem ser encarados como homúnculos, pequenas partes a que se atribuem uma parcela de intencionalidade. Conforme o que foi explicado anteriormente, para que não se incorra na falácia do homúnculo é necessário que todos esses subsistemas diminutos possam ser efetivamente compensados através de níveis de projeto que não pressuponham a intencionalidade.

Essa tarefa é realizada da seguinte maneira: . (DENNETT 2006 p 62)divide-se os homúnculos em partes cada vez menores, sendo que cada uma das partes é um tipo de verificador ou avaliador com menos informações do que os sistemas de nível mais elevado, ou seja, com *menos tarefas definidas intencionalmente.* Prosseguindo nessa cadeia sucessiva de subdivisões em partes menores, os pesquisadores finalmente chegam aos níveis mais baixos possíveis, níveis esses em que somente os princípios elementares da física atuam, sem que se pressuponha nenhuma intencionalidade. Pode-se atingir um entendimento mecanicista compreensível dos processos envolvidos no funcionamento de um sistema de IA. Desse modo:

> O programador de IA utiliza a linguagem intencional sem receios porque ele sabe que se for bem sucedido em fazer seu programa funcionar, quaisquer pressuposições circulares que ele possa ter feito vão, provisoriamente, ser desfeitas. Se o programa funcionar, então podemos estar certos que os homúnculos foram eliminados da teoria. (DENNETT 2006 p 62)

A tarefa de atribuir a intencionalidade às máquinas depende de saber se podemos, em princípio, formular uma explicação da intencionalidade que não

seja circular, ou seja, a teoria não pode pressupor a intencionalidade nos níveis mais básicos dos sistemas a serem considerados. Uma teoria da intencionalidade deve conter postulados os mais simples possíveis, sem permitir a existência tautológica de homúnculos não explicados.

2.1IA e tese de Church

Conforme o exposto acima, pode-se mencionar a tese de Church. Uma vez que essa tese estabelece que qualquer procedimento algoritmo possível deve se fundamentar nos limites do computável, vemos que uma teoria satisfatória da intencionalidade seria aquela cujos elementos básicos se restrinjam às funções computáveis, funções que podem ser executadas por procedimentos que não pressuponham a intencionalidade. Podemos considerar que a postura intencional é compatível com a tese de Church, uma vez que a atribuição de racionalidade a um sistema é compensada devido ao procedimento de se reduzir o sistema em causa aos níveis mais fundamentais possíveis, níveis simples o bastante para serem executados por procedimentos mecânicos.

As afirmações feitas acima se aplicam também na comparação entre a pesquisa de IA e teorias dualistas e antimecanicistas da intencionalidade. Uma vez que a IA é o estudo dos modelos mecanicistas da inteligência, pode-se ver que, ao fazer a redução última ao nível de linguagem de máquina, a IA está de acordo com os limites da computabilidade estabelecidos pela tese de Church, eliminando quaisquer possíveis homúnculos não explicados na teoria. Em contrapartida, o dualismo postula que a intencionalidade possui aspectos não físicos, não redutíveis aos parâmetros das funções computáveis. E, uma vez que o dualismo não se limita aos procedimentos computáveis simples o bastante para serem realizados por um algoritmo mecânico desprovido de intencionalidade, os aspectos não físicos postulados por uma teoria dualista são, eles próprios, homúnculos não eliminados que de maneira circular pressupõem a intencionalidade, fazendo com que uma teoria dualista não seja satisfatória. Por comparação, a pesquisa em IA, devido ao fato de reduzir a pressuposição de intencionalidade a um programa mecanicista, se justifica enquanto teoria não tautológica.

2.2 Argumentos contra a Turing-computabilidade

Pode-se questionar se a intencionalidade das máquinas seria efetivamente verdadeira. Esse questionamento também se baseia na acepção de que a intencionalidade pode ser dividida em duas categorias distintas: intencionalidade original e derivada. (SEARLE,1995). Segundo essas acepção, a intencionalidade da mente humana é original, enquanto os artefatos mecânicos derivariam sua intencionalidade da mente que os projetou. Pretende-se então concluir que a intencionalidade de um sistema de IA não seria genuína, mas sim um mero reflexo dos objetivos dos engenheiros que projetaram o sistema. Ocorre que, enquanto seres biológicos, os engenheiros derivariam a intencionalidade de suas mentes ou cérebros do processo não intencional da seleção natural:

> [...] os símbolos presentes em nossas mentes derivam sua intencionalidade da economia que desempenham no sistema funcional cerebral de que tais símbolos fazem parte, sendo que esse sistema foi desenvolvido pelo processo de evolução por seleção natural. (DENETT, 1996, p 82)

Assim como obtemos nossa intencionalidade do funcionamento cerebral, em última análise devido ao processo não randômico de seleção genética, analogamente um robô, ainda que construído e programado deliberadamente, obteria sua intencionalidade das funções de suas subrotinas. As sub-rotinas operaram de acordo com as regras da física, regras que, no dizer de Hofstadter " não possuem nenhuma meta na frente" (HOFSTADTER, 1979, p. 683).

De maneira análoga ás atividades do sistema nervos humano, um sistema de IA pode apresentar uma genuína intencionalidade devido apenas á interação de seus sub-programas, que em seu nível mais elementar, atuam segundo princípios físicos, princípios que não necessitam das regras programadas pelos pesquisadores.

CAPÍTULO 3: O TEOREMA DE GÖDEL E O QUARTO CHINÊS

No contexto da filosofia da mente há uma argumentação teórica que utiliza o teorema de Gödel para corroborar a alegação de que a inteligência artificial é impossível, ou seja, as máquinas não podem apresentar uma forma de intencionalidade. Alega-se que a mente humana não é um sistema formal de manipulação de símbolos tal qual um computador eletrônico.

Basicamente o teorema de Gödel, proposto pelo matemático Kurt Gödel (Gödel,1962) diz respeito a um problema que ocupou os matemáticos e filósofos no início do século XX, a saber, se era possível derivar toda a matemática de um conjunto fixo princípios formalizados. Dito de outro modo: procurava-se saber se seria possível provar qualquer teorema a partir de alguns poucos axiomas matemáticos. O teorema matemático de Gödel estabelece que essa sistematização não pode ser alcançada. Há uma incompletude em qualquer axiomatização da matemática e isso é determinado pelos teoremas de Gödel:

Teorema 1) Uma teoria axiomática qualquer enumerável de modo recursivo e capaz de expressar verdades aritméticas não pode ser consistente e completa. Dada uma teoria consistente qualquer haverá proposições que não podem ser refutadas ou demonstradas.

Teorema 2) Uma teoria capaz de expressar verdades aritméticas só pode provar sua própria consistência apenas se tal teoria for inconsistente .

Segundo o primeiro teorema da incompletude de Gödel, certas proposições de um dado sistema são indecidíveis, pois não podem ser provadas ou refutadas através de axiomas do sistema em causa.

1. Máquinas de Turing e teorema de Gödel

O uso dos teoremas da incompletude para argumentar contra o programa de inteligência artificial foi inaugurado pelo próprio Gödel. Em conversas preservadas pelo lógico Hao Wang, Gödel afirma:

> 6.12 Ou a mente humana supera todas as máquinas(para ser mais preciso: pode-se decidir mais questões numérico-teóricas que qualquer máquina), ou então existem questões numérico-teóricas indecidíveis para a mente humana[não se exclui que as duas alternativas possam ser verdadeiras.] (Wang, 1997,pg.185)
>
> 6.14 Ou a matemática subjetiva supera a capacidade de todos os computadores ou então a matemática objetiva supera a matemática objetiva ou as duas alternativas são verdadeiras.
>
> 6.15 Se a primeira alternativa se sustenta, então parece implicar que o funcionamento da mente humana não pode ser reduzido ao funcionamento do cérebro, que parece ser uma máquina finita com um número finito de partes, chamadas neurônios e suas conexões.
>
> 6.18 Meu teorema da incompletude mostra ou que a mente humana não é mecânica ou que a mente não pode entender seu próprio mecanismo.(Wang,1997. Pg 186

Segundo Gödel, a mente ou o cérebro humano não funciona de maneira puramente algorítmica. o que significa que a cognição humana não pode ser entendida como um sistema de processamento simbólico de informações. A abordagem computacionalista ou cognitivista, associada a filósofos contemporâneos como Dennett e Fodor, postula que os estados mentais são estados simbólicos associados em uma seqüência de outros símbolos, sendo que os processos mentais são transformações ou manipulações lógicas dessas

cadeias simbólicas, de modo análogo às operações de uma máquina de Turing[1].

Suponhamos, de acordo com Dennett, que determinado ser humano fosse um tipo de máquina de Turing TMg, então haveria uma sentença de Gödel sobre TMg que tal humano não poderia provar. Ocorre que Gödel alega que o humano em questão pode de fato provar a referida sentença de Gödel e, assim sendo, conclui-se que nenhum ser humano é uma máquina, permanecendo além de qualquer análise computacional passível de ser aplicada pela ciência. Ocorre que, segundo Dennett:

> Desejo mostrar que todos esses argumentos devem falhar porque, em um ponto ou outro, devem implicitamente negar uma verdade óbvia, a saber, que as exigências da lógica exercem sua força não sobre as coisas do mundo diretamente, mas antes, sobre o que devemos considerar descrições ou interpretações defensáveis das coisas. (DENNETT, 2006, p. 338).

Pode-se dizer que os críticos da IA como Gödel erram ao suporem que a determinação das ações a habilidades de um homem, e também a determinação das habilidades de um processador de informações, podem prosseguir de uma maneira que não resultará em um raciocínio tautológico sobre a aplicabilidade do teorema da incompletude de Gödel. Dito de outro modo, o erro dos céticos se deve ao fato de que, conforme será mostrado ao longo do presente capítulo, diferentes interpretações de máquina de Turing são aplicáveis a qualquer homem ou dispositivo. Assim sendo, um ser humano ou computador pode produzir uma demonstração da sentença de Gödel através de uma interpretação de máquina de Turing específica, apesar de que a incompletude sempre é restabelecida. (DENNETT, 2006, p. 340).

O teorema de Gödel versa sobre a inconsistência de um sistema matemático, onde há enunciados que não podem ser provados com base em

---

[1] JOHNSON-LAIRD, P. *The Computer and the Mind*. Cambridge : Havard University Press, 1989.

axiomas desse sistema (NAGEL; NEWMAN, 1958). Uma máquina de Turing - também chamada de computador universal- é um sistema formal de cálculo que, ao ser estabelecido pelo matemático britânico Alan Turing, lançou as bases fundamentais da computabilidade. As máquinas de Turing encerram um conjunto de diretrizes, correspondentes a regras de inferência de procedimentos axiomáticos formalizados. Pode-se compreender uma máquina de Turing como sendo basicamente o conjunto dos aspectos lógicos de qualquer computador, englobando noções como capacidade de memória, estados e transições. Uma vez que as máquinas de Turing são conceitos abstratos tanto quanto um sistema de axiomas, o teorema de Gödel se refere a tais máquinas.

Uma máquina de Turing consiste em nada mais que um conjunto de instruções que realizam operações sobre seqüências de símbolos. As instruções por sua vez são coligidas na forma de estados de máquina, sendo que cada um desses estados é, ele mesmo, um conjunto de instruções. Há adicionalmente uma função de mudança de estado que prescreve qual seqüência de instruções deve ser seguida em resposta ao dado de entrada**.** Essa função de mudança de estado ou função de transição orienta a máquina sobre que símbolo deve ser escrito e como o discriminador de símbolos deve se mover, considerando os símbolos escritos em uma fita. Há ainda um registrador de estados, que tem a função de registrar o estado da máquina de Turing. Essa definição de máquina de Turing é completamente abstrata no que se refere ao modo como tal operação deve ser realizada, desse modo uma dada máquina de Turing pode ser realizada de diversas maneiras diferentes, tais como: um conjunto de pessoas podem constituir uma máquina de Turing, realizando cada uma delas operações em cartões onde estão escritas as instruções de mudança de estado de máquina; outro possível exemplo seria um aparelho mecânico que realiza leitura de fita analógica. Por fim , podemos citar a usual simulação em computador digital.

O processo de construção de uma máquina de Turing específica envolve, a princípio, o alcance das regularidades especificadas pela máquina de Turing em questão.. Isso significa que para cada símbolo através do qual o

input e o *output* serão expressos, o dispositivo que realizará a máquina de Turing deve ser construído de uma maneira mecanicamente distinta, como um furo específico em uma fita ou mesmo um caractere marcado em um cartão de papel. Por razões de parcimônia, esses detalhes devem apresentar uma certa estabilidade, ser executáveis com uma determinada rapidez e também apresentar dimensões reduzidas. Desse modo, temos um projeto o mais próximo possível de uma máquina ideal. Cumpre notar que os símbolos devem ser discriminados por dispositivos compatíveis com a simbologia adotada. Não importando qual a devida natureza da reação desses dispositivos que são capazes de ler os símbolos, eles devem ser construídos de forma que cada qual seja diferente um do outro, possibilitando que cada seqüência de dispositivos possa reagir de forma diferenciada a uma cadeia consecutiva de símbolos. Na máquina de Turing, cada mudança de estado de máquina deve ser seguida de maneira correlacionada por uma mudança física, sendo que essa mudança física correspondente pode ser o deslocamento de um conjunto de polias, o fechamento de um circuito elétrico ou algum movimento produzido por um tubo pneumático (DENNETT, 2006, p. 341)..

Um aspecto de suma importância é que a máquina deve ser projetada com o máximo de isolamento possível do meio ambiente externo, de modo a evitar que mudanças de temperatura ou mesmo umidade interfiram no funcionamento completo da máquina. O que diferencia uma mudança física interpretada como *input* ou uma mudança causada por interferência de condições externas, em síntese, o que nos permitir distinguir entre as operações da máquina de Turing e os fenômenos do entorno físico em que a máquina se encontra em determinado momento no tempo, são as leis físicas que foram aplicadas no projeto da máquina. Desse modo, um furo em uma fita pode ser um símbolo em uma máquina ou um acontecimento que cause defeito em outra máquina. Esse raciocínio se aplica também às mudanças internas da máquina. Algumas mudanças internas serão mudanças de estado de máquina ou falhas, dependendo da economia de engenharia do projeto adotado. Caso, por exemplo, encontremos uma máquina que seja afetada por mudanças de temperatura no ambiente, poderemos concluir que embora para muitas

aplicações tal máquina não apresente um bom desempenho, é possível que ela possa ser utilizada como um termostato.

Se hipoteticamente o projeto do aspecto físico da máquina for algo bem estabelecido, e a finalidade de uma máquina puder ser conhecida com relativa facilidade, poderemos desconsiderar a distinção acima mencionada entre os detalhes de engenharia e o meio externo. Com algum grau de exatidão poderemos observar de modo direto o *input*, o *output* e as mudanças de estado de físico para determinarmos com exatidão as funções da máquina de Turing projetada. Contudo, essa identificação se mostra de fato praticamente além dos limites do que pode ser executado. Para ilustrar esse fato, imaginemos.. dois estudiosos distintos, A e B, que ao se depararem com um componente de maquinaria se deslocando em cima de uma fita, procuram ambos estudar as atividades da máquina através do tempo, desmontando a máquina e observando cada componente particular separadamente, tornando a remontar o dispositivo para tentar compreender qual a natureza da máquina em causa.

De acordo com Dennett (DENNETT, 2006, p.341), poderíamos suspeitar que o desacordo entre A e B seria apenas a respeito do que consiste em interpretação de símbolos de *input* e *output*. Assim sendo, a discordância seria em relação a finalidade da máquina. Tanto A como B não saberiam encontrar um ponto pacífico a respeito do objetivo para o qual a máquina foi projetada. Essa discordância faria com que A trate a simbologia da máquina de Turing como números, e através dessa abordagem considerar que a máquina consista em um dispositivo de determinação de números primos. O estudioso B, por sua vez, trata a simbologia da máquina como elementos operadores de alguma linguagem, e com base nisso chega a conclusão que a máquina de Turing é um dispositivo de prova de teoremas matemáticos, um dispositivo capaz de captar alguns eventos do mundo exterior, processá-los e em seguida gerar seqüências de fórmulas interpretáveis como provas de teoremas. O foco central do desacordo entre A e B é apenas em relação ao significado das regras e funções da máquina. A natureza abstrata e formal de uma máquina de Turing significa que teoricamente tal máquina pode estar cumprindo qualquer

finalidade possível, de acordo como sua simbologia é interpretada semanticamente.

Suponhamos que o estudioso A faça a suposição de que a máquina em questão determina os logaritmos dos números que lhe são fornecidos como *input*, mas apesar disso, A consiga mostrar que a máquina apresenta, a despeito de todos os detalhes de engenharia, um defeito no sentido de fornecer resultados falsos para determinadas seqüências de números (digamos, para fins de argumentação, que a máquina erre ao determinar qualquer logaritmo na base 10). É importante ter em mente que A aponta esse defeito unicamente com base na suposição que faz sobre o que a máquina de Turing foi projetada para fazer, e não com fundamentação em detalhes fornecidos pelos construtores da máquina. Se , além disso, A disser que a máquina não é adequadamente imune a mais moderada interferência externa, pois qualquer choque mecânico pode fazer com que os discriminadores de símbolos realizem uma interpretação ineficaz dos dados, prejudicando com isso o resultado dos dados de saída. De acordo com Dennett,(DENNETT, 2006, p.342) não obstante essas colocações de A, o estudioso B alega que a máquina em questão consiste em um aparelho capaz de identificar padrões nas marcas da fita lida pelos discriminadores de símbolos. Os dados de saída são elementos de alguma linguagem artificial, algo remotamente análogo a linguagem de programação LISP, sendo que essa linguagem descreve os padrões detectados na fita pelos discriminadores. A máquina, no dizer de B, ajusta suas mudanças de estado em correlação adequada às condições exteriores. Para o estudioso B, os dados de saída são mal gerados apenas devido a um pequeno componente incorretamente colocado na máquina, o que ocasiona o mencionado processamento inadequado dos dados de saída. As interpretações e suposições de A e B não são dessa vez apenas com relação a finalidade da máquina, pois nesse caso:

> Há um desacordo não apenas sobre o propósito da máquina, ou a semântica da linguagem que ela utiliza, mas também sobre a sintaxe e o alfabeto. Não há uma correspondência biunívoca entre suas enumerações de símbolos e instruções. (DENNETT, 2006, p.342)

O fato em questão é a discordância entre os detalhes da máquina que constituem aspectos de engenharia e, por outro lado, daquilo que nada mais seria do que apenas falhas de projeto. Para melhor demonstrar esse ponto da argumentação, pode-se dizer que tanto A como B partem da mesma descrição dos estados físicos da máquina para fazer em seguida a mesma predição das ações da máquina ao longo do tempo. Mas não existe nesse caso um acordo entre quais partes da economia interna da máquina podem ser consideradas como partes de projeto e quais detalhes são nada mais que defeitos. O desacordo entre os dois estudiosos também se aplica a quais dados de saída devem contar como parte da simbologia da máquina de Turing.

Um corolário interessante da argumentação em causa são os desacordos que podem surgir entre A e B em relação a dois aspectos distintos na interpretação da máquina:

1) Pode-se tratar qualquer evento como um possível *input* ou dado de entrada;

2) Não existe uma única interpretação de máquinas de Turing; disso decorre o fato de que se um determinado dispositivo for interpretado com uma realização concreta de uma máquina de Turing Tx, é igualmente possível interpretar esse mesmo dispositivo como uma realização de uma máquina de Turing Ty.

Cada um desses itens enumerados acima podem ser explicados de maneira mais detalhada. Primeiramente, quanto aos dados de *input,* mesmo que em um primeiro momento a máquina não apresente nenhuma reação aos eventos do mundo externo, não podemos estar completamente seguros aprioristicamente que não haverá reação alguma em um momento futuro qualquer. Além disso, não ocorre uma delimitação física claramente perceptível a princípio entre quais aspectos são partes constituintes da máquina e quais aspectos não fazem parte das operações realizadas pela máquina.

Quanto às diversas interpretações de máquinas de Turing, pode-se dizer,de acordo com o Dennett, que embora um computador pessoal possa ser

visto como uma realização de uma máquina capaz de exibir informações escritas em linguagem natural, ele também pode de modo adequado ser visto como uma realização de uma máquina capaz de exibir imagens e vídeos.

A partir do momento em que nos deparamos com as interpretações diferentes propostas por A e B (DENNETT, 2006, p. 343)., pode-se dizer que a melhor forma de determinar qual interpretação é a mais adequada não consiste apenas em pressuposições sobre o que consideramos mais plausível no que se refere às operações da máquina, mas deveríamos procurar saber quais os objetivos intencionais dos engenheiros ao projetar a máquina. Procurar descobrir as intenções dos projetistas a respeito de qual máquina de Turing eles estavam tentando realizar não revelaria, contudo, uma forma de decidir objetivamente qual a interpretação correta do dispositivo que os estudiosos A e B estão analisando. Se for possível interpretar o dispositivo como um calculador de logaritmos, então, segundo essa interpretação, o dispositivo é um calculador de logaritmos independentemente do caso de um grupo de engenheiros haverem intencionado construir uma máquina capaz de realizar outra função específica. Analogamente, se concluirmos que uma interpretação aplicável ao dispositivo é tratá-lo como um identificador de padrões, então ele pode ser considerado um identificador de padrões tão eficiente quanto um aparelho hipotético que fosse construído com essa finalidade. O cerne dessa digressão é que para qualquer objeto possível há como considerá-lo como uma realização concreta das mais diferentes formas de máquinas de Turing, sendo que caso escolhemos uma forma específica, então o objeto em questão pode ser visto como um tipo de máquina de Turing de acordo com essa interpretação escolhida, e essa forma de interpretação não exclui as outras realizações de máquina de Turing que porventura podemos aplicar ao objeto hipotético.

No conjunto de objetos possíveis, há os organismos biológicos; sendo que é possível interpretar as funções de um organismo como operações de uma máquina de Turing. Além disso, o sistema nervoso realiza funções equivalentes a uma forma de computação, gerando padrões de saída após o processamento de padrões de entrada e depois de consulta em informações

armazenadas nas áreas nervosas responsáveis pela memória. No dizer de Dennett:

> Podemos supor que um animal pode com proveito ser visto como um computador ou um autômato, e uma vez que qualquer autômato pode ser simulado por uma máquina de Turing, isso equivale de certo modo à suposição de que poderíamos querer tratar um animal como uma máquina de Turing. (DENNETT, 2006, p. 343).

A dúvida que decorre (DENNETT, 2006, p. 343) consiste em saber se podemos ou não decidir em princípio qual é a máquina de Turing correta para a interpretação aplicada às formas de vida biológicas existentes . Assim como no exemplo do dispositivo mecânico de leitura de fita, ao interpretarmos um ser vivo como uma máquina de Turing, precisamos decidir primeiramente quais os efeitos causados sobre o ser vivo contam como dados de entrada e quais efeitos são mera interferência, e não há maneiras *a priori* de decidir essa questão. Mesmo se decidirmos utilizar critérios referentes à temperatura ou pressão e , com base nisso, perguntarmos se em relação àquele ser vivo em causa as mudanças de temperatura ou pressão constituem dados de entrada ou interferência, podemos descobrir que em algumas situações ocorrem mudanças consideráveis mediante a certos níveis de temperatura e pressão, mas ainda assim não podemos estar completamente seguros se essas mudanças devem se analisadas como dados de *input* ou se, ao contrário, para uma interpretação de máquina de Turing específica essas mudanças quantificáveis no ser vivo não apresentam nenhum significado relevante.

Uma semelhança entre a procura por uma definição de máquina de Turing aplicável ao ser vivo e ao dispositivo mecânico é a pressuposição de que uma máquina de Turing sempre deve ter algum propósito ou finalidade. Em um sentido lógico, essa pressuposição não é absolutamente segura. Isso se deve porque, segundo a tese de Church (essa tese consiste na acepção de que qualquer algoritmo pode ser implementado por uma máquina de Turing), qualquer função computável pode ser processada por uma máquina de Turing. O significado disso é que uma máquina de Turing pode computar as mais diversas funções computáveis(embora precise ser reiterado que a máquina de

Turing só pode computar funções computáveis. Há funções, como a função da parada que não são computáveis. Dito de outro modo: não existe um procedimento algorítmico que possa decidir corretamente se proposições matemáticas arbitrárias são verdadeiras ou falsas), ainda que uma função específica não tenha nenhuma finalidade. O fato de estar computando uma função sem propósito ou objetivo não é uma razão para que a máquina não atenda a todos os critérios necessários para ser uma máquina de Turing.

De um ponto de vista meramente científico, tratar um objeto artificial como uma máquina de Turing apresenta relevância conceitual para fins de pesquisa, se pudermos considerar aquelas partes do objeto que podem ser utilizadas para uma finalidade ou objetivo. E, no caso do estudo dos seres biológicos, o conceito de máquina de Turing é importante se, e somente se, for possível realizar um estudo das pares funcionais adaptativas dos organismos biológicos.

Ao aplicarmos uma interpretação de máquina de Turing a um animal biológico podemos descobrir que, em certas situações, as mudanças de pressão possuem um significado como dados de *input*, fazendo com que o animal se movimente de forma satisfatoriamente adaptativa, procurando abrigo ou refúgio, por exemplo. Desse modo, ao procurarmos uma explicação para a movimentação do animal, estaremos certos em tratá-lo como uma máquina de Turing, sendo que os efeitos de pressão sobre o animal constituem dados de entrada ou sinais de informação que são processados pelo sistema nervoso, gerando os dados de saída que nada mais são do que o movimento observado.

Ainda em relação ao tratamento de um animal específico como uma máquina de Turing (DENNETT, 2006, p. 345), precisamos decidir quais aspectos de sua fisiologia consistem em aspectos adaptativos no ambiente natural do animal em causa, e quais aspectos não são funcionalmente adequados no organismo do animal. O comportamento de um cachorro que, ao escutar o som de um sino, começa a salivar, não constitui um comportamento para o qual o cachorro está naturalmente propenso. Dizemos nesse caso, que se trata de um comportamento estimulado adquirido por um processo deliberado de reflexo condicionado. Ainda que possamos descrever o

comportamento do cachorro no caso em questão como uma máquina de Turing, onde o ruído do sino é o dado de *input* e o correspondente ato de salivar é o dado de saída, podemos estar certos que esse comportamento não foi desenvolvido pela evolução como uma resposta naturalmente adaptativa. Qualquer objeto, seja um animal ou um artefato, pode ser descrito como diferentes formas de máquinas de Turing, de acordo com o que consideramos como mudanças de estado de máquina ou dados de entrada. Disso decorre que não dispomos de meios para decidir o que deve contar como a interpretação de máquina de Turing correta. Como dito anteriormente, a discussão sobre máquinas de Turing apenas possui relevância conceitual se fizermos considerações sobre os possíveis aspectos funcionais do animal, e se não conseguirmos decidir a respeito das funções de partes da fisiologia do ser vivo em causa, não poderemos em conseqüência alcançar uma compreensão evolutiva do animal.

Podemos tomar como exemplo uma forma de vida humana e perguntar se um ser humano é uma máquina de Turing(DENNETT, 2006, p. 345). É fato que o homem pode ser visto como um objeto demasiado complexo do mundo empírico. E, uma vez que o homem é um objeto material dotado de partes mecânicas, ele certamente pode ser descrito como as mais variadas realizações de máquinas de Turing possíveis. Ocorre que, dentre as mais diversas interpretações de máquinas de Turing, existem aquelas formas que podemos dizer que apresentam sentido matemático. Suponhamos por exemplo, que o feixe de luz visível que atravessa a retina no sistema visual humano possa ser interpretado como um dado de *input* e, além disso, interpretamos o movimento dos braços como uma forma de *output* ou dado de saída. Feitas essas pressuposições, consideremos o referido movimento de braços como uma maneira de solução ou provas de teoremas de uma geometria descritiva. Essa interpretação matemática parece contraintuitiva na medida em que acreditamos que um simples movimento de braços não tem por finalidade produzir provas matemáticas. Todavia, ao mudarmos nossas pressuposições sobre o que denominar como *input*, podemos de fato interpretar qualquer humano como todas as máquinas de Turing possíveis. Isso significa que mesmo a movimentação de membros do corpo ou mesmo

qualquer ação humana pode ser vista como maneiras de produzir provas de teoremas de qualquer geometria, ainda que, de acordo com Dennett:

> [...] as noções que constituiriam esses feitos de prova não *se pareceriam com* feitos de prova, mas com dormir, comer, falar sobre o tempo. O antimecanicista não está interessado em interpretações de máquinas de Turing desse tipo; as atividades e habilidades sobre as quais ele pressupõe ter informação crucial são aquelas de matemáticos em seus empreendimentos profissionais. (DENNETT, 2006, p. 345)

Gödel e os críticos da IA , ao tentarem demonstrar que a intencionalidade das máquinas é impossível, estão procurando uma definição de humano cujas ações e objetivos interpretáveis sejam evidentes. Essa definição supostamente se basearia na assertiva de que a mente não pode ser encarada como um objeto puramente mecânico. Segundo Lucas:

> [...] tenta-se produzir um modelo da mente mecânico – essencialmente morto- embora a mente, devido ao fato de ser viva, sempre é superior a qualquer sistema formal e morto. Devido ao fato de estar viva, a mente sempre tem a ultima palavra. (apud HOFSTADTER,1979, p.470)

Ocorre que,, uma vez que perguntemos qual interpretação de máquina de Turing se ajusta adequadamente a essas ações e, em conseqüência, qual interpretação poderia ser a única correta para ser aplicada ao um ser humano, voltaremos para o problema discutido anteriormente, a saber: existem múltiplas interpretações de máquinas de Turing passíveis de serem aplicadas a um objeto, como por exemplo, um ser humano. Se a hipotética interpretação correta for aquela que apreende satisfatoriamente um ser humano enquanto dispositivo funcional biológico, e visto qualquer humano tem ações diferentes além de provar teoremas, podemos perceber que nenhuma forma de interpretação de máquina de Turing seria a única correta para estar de acordo com os argumentos levantados pelos críticos da IA. Pode-se estar certo de que:

> [...] além de toda computação que um homem possa fazer (na escola, nos negócios, para se divertir), ele também come, consegue abrigo, faz amigos, protege-se, e assim por diante; não precisamos de Gödel para mostrar que o homem não é apenas um computador *neste* sentido-- isto é, um dispositivo cujo único propósito é o de computar funções ou provar teoremas. (DENNETT, 2006, p. 346).

Podemos imaginar como exemplo um tipo de máquina de Turing TGa que seja especificada como tendo a função de produzir demonstrações de teoremas (DENNETT, 2006, p, 347). Vamos supor que um determinado aparelho de computador seja uma realização de TGa, demonstrando teoremas com uma certa regularidade, através de um sistema de axiomas. Além disso, consideremos um matemático humano que esteja, ele próprio, demonstrando os mesmos teoremas com a mesma base axiomática. Pode-se dizer que, assim como o computador, o matemático é uma realização concreta da mesma máquina de Turing TGa. Assim sendo, o teorema de Gödel se aplica a ambas as realizações da máquina de Turing exemplificada. Podemos estabelecer os movimentos do matemático que constituem os dados de saída da máquina de Turing mencionada, sendo que esses movimentos estarão correlacionados às seqüências de símbolos exibidos pelo computador. Se em determinado momento no tempo for solicitado ao matemático que execute outra atividade, suspendendo apenas por um instante seus cálculos de demonstrações de teoremas, ocorrerá uma mudança nos dados de saída. O aspecto central do exemplo proposto consiste em que essa suspensão dos cálculos não mostra que o matemático humano não é uma máquina ; a única coisa que fica estabelecida decisivamente é que o matemático (e, por extensão, os humanos em geral) não é somente um tipo de máquina, no caso em questão, uma realização de máquina de Turing TGa. Ao interromper os cálculos ocorre uma alteração na seqüência da simbologia dos dados de saída, alteração que é contrária às instruções para a implementação da máquina de Turing TGa. E, uma vez que o seja possível que as ações do matemático sejam contrárias às instruções de realização de uma TGa, fica-se estabelecido que ele não é apenas uma realização momentânea de uma máquina de demonstração de teoremas.

Pode-se dizer que o matemático é uma realização de uma máquina TGa devido ao fato de que, por um momento, ele simulou uma TGa seguindo as devidas instruções necessárias. Uma simulação é, nesse sentido, uma realização de uma máquina de Turing. Ocorre que, de acordo com o argumento dos céticos da IA, se um homem fosse uma realização de uma máquina TGa, ele não poderia provar uma sentença S (S, nesse caso, é uma sentença de Gödel sobre TGa). Para Gödel, só é possível provar a sentença S porque homens não são nenhum tipo de máquina de Turing. Contudo, se ser ou realizar uma máquina TGa consiste em simular uma TGa, então podemos mostrar que um ser humano pode provar a sentença de Gödel sobre TGa. É necessário somente que o humano não execute as instruções correspondentes à simulação de uma TGa. Isso se aplica também ao aparelho de computador mencionado acima: é possível para o aparelho provar a sentença S sobre TGa se, e somente se, o computador utilizar outra base ou sistema axiomático diferente do sistema referente a uma execução de uma máquina TGa (existem muitos programas de IA que podem ampliar a sua base de axiomas). Para qualquer instanciação concreta de máquinas de Turing essa mudança de instruções é possível. O teorema de Gödel somente estabelece como uma limitação a impossibilidade de provar a sentença S de TGa enquanto as devidas instruções necessárias a uma realização de uma TGa estiverem sendo adequadamente executadas. Essa limitação gödeliana se aplica a qualquer objeto possível, seja esse referido objeto um ser humano biológico ou até mesmo um sistema artificial de computador. A réplica mais direta possível ao argumento do teorema de Gödel é que, embora esteja estabelecido que há limitações às capacidades de qualquer máquina, enunciou-se, sem nenhuma espécie de prova, que essa limitação não se aplica à inteligência humana. (TURING,1950).

2. O Argumento do Quarto Chinês

Além do argumento do teorema de Gödel, uma das críticas mais discutidas em relação à inteligência artificial consiste no conhecido argumento do quarto chinês, proposto por John Searle[2]. Esse argumento procura criticar a asserção de que o cérebro humano e um computador sejam de algum modo análogos, não somente quantitativa mas também qualitativamente. Assim como o argumento do teorema de Gödel, o quarto chinês procura defender a hipótese de que computadores não podem exibir intencionalidade. A concepção de que um cérebro é um computador digital foi chamada por Searle de Inteligência Artificial Forte ou simplesmente IA Forte. Essa concepção sustenta que o cérebro é uma máquina enquanto que a mente é um programa que está sendo rodado no cérebro. Para a comunidade científica envolvida com essa acepção, o fenômeno da inteligência, seja natural ou artificial, consiste em uma questão de manipulação simbólica formal, e não um fenômeno relacionado às propriedades físico-quimicas essenciais dos objetos. Searle pretende mostrar com seu argumento filosófico que a concepção funcionalista, defendida pelos pesquisadores de IA, está equivocada ao supor que podemos projetar uma forma de inteligência apenas construindo algo que instancie um programa adequado de manipulação de símbolos, não importando a natureza intrínseca do material que porventura utilizemos. Esse programa hipotético poderia, em princípio, ser elaborado com diversos materiais. Seus componentes estruturais poderiam ser construídos de compostos orgânicos ou até mesmo circuitos integrados de materiais semicondutores fabricados principalmente com elemento químico silício. Um programa de computador, de acordo com os pesquisadores de IA, poderia apresentar pensamentos e sensações emocionais comparáveis aos pensamentos e sentimentos humanos, unicamente em vista de instanciar um programa estruturado adequadamente. Em síntese, a pesquisa de IA estabelece que qualquer sistema físico organizado de uma maneira funcionalmente apropriada para o processamento de símbolos pode em tese, ser inteligente. A operação de um programa com

[2] SEARLE,J. " Minds, Brains and Programs" In: *Behavioral and Brain Sciences* 3, 1980, p. 417-424.

dados de entrada e saída, não importando a constituição de seus componentes, é suficiente para ser o equivalente à inteligência natural humana. Para Searle, uma mente humana capaz de apresentar intencionalidade é um fenômeno essencialmente biológico subjetivo irredutível, não podendo ser compreendido apenas como uma atividade reproduzível por um programa rodado em um artefato constituído por um material distinto da constituição do cérebro orgânico.

Antes de apresentarmos o argumento propriamente dito, é necessário discorrermos a respeito de alguns pressupostos envolvidos argumentação de Searle. Isso será necessário para compreendermos as críticas feitas por Dennett com o objetivo de refutar o argumento do quarto chinês. Basicamente, Searle alega que o centro da questão não consiste no grau de complexidade de um dispositivo computacional, mas antes toda a sua crítica da posição de que o cérebro é um computador digital está baseada na própria natureza fundamental do funcionamento de um computador. Um computador construído com qualquer tecnologia razoavelmente avançada apresenta operações essencialmente formais. Suas operações de processamento simbólico de informação são definidas especificamente em termos de conjuntos de seqüências binárias de zeros e uns, que, por sua vez, são impressos digitalmente. Uma possível regra de mudança de estado de máquina embutida no computador prescreve que, quando ocorre determinado processo na programação, um símbolo deve ser impresso ou mesmo removido. O processamento de um novo símbolo ou simplesmente o apagamento de um símbolo constituem formas de execução de um novo estado de máquina. Esse estado será seguido de alguma mudança física, conforme a anteriormente mencionada definição de máquinas de Turing. Ocorre que, de acordo com Searle:

> Os símbolos não têm significado; não têm conteúdo semântico; não são acerca de qualquer coisa. Têm de ser especificados unicamente em termos de sua estrutura formal ou sintática. Os zeros e os uns por exemplo, são simples numerais. (SEARLE, 1984, p. 30)

A característica fundamental de um programa formal, de acordo com Searle, é que qualquer programa não apresenta entendimento sobre o que quer que seja. Ainda que se possa projetar um aparelho que possa processar qualquer programa escrito em qualquer linguagem de programação, um programa permanece sendo, de acordo com sua própria definição, um conjunto de instruções formais abstratas de manipulação sintática de símbolos. Uma vez que segundo Searle, os estados mentais apresentam intencionalidade e consciência em virtude de possuírem conteúdo irredutível à simples operações sintáticas, os estados que constituem uma mente inteligente não podem ser semelhantes a um programa de processamento sintático de informação. Dito de outro modo: Uma vez que um estado mental apresenta intencionalidade (um estado mental é sobre alguma coisa no mundo), todo estado mental apresenta um conteúdo diferente de operações de sintaxe formalmente definidas. Se uma cadeia de seqüências sintáticas não apresenta, por si só, nenhum significado ou intencionalidade, deve haver, segundo esse raciocínio algo que torna possível o conteúdo de um estado mental. Esse algo consiste, segundo Searle, na semântica intimamente correlacionada com os estados mentais do cérebro. Um programa de computador não pode possuir uma intencionalidade autêntica porque, de acordo com Searle, um programa nada mais é do que mera sintaxe abstrata. Um programa não apresenta, e nem mesmo é possível que apresente em princípio, um conteúdo semântico semelhante ao conteúdo da mente humana.

Se o raciocínio de Searle for colocado em uma estrutura lógica conceitual, ele apresentará a seguinte forma, a saber:

1. programas são sintáticos;

2. mentes têm conteúdo semântico;

3. A sintaxe não é suficiente para a semântica;

4. implementar um programa não é suficiente para se obter uma mente;

Para exemplificar esse raciocínio lógico, Searle propôs um experimento filosófico conhecido como argumento do quarto chinês. O argumento pode ser visto como uma resposta crítica ao assim chamado teste de Turing. Esse teste procura estabelecer se uma máquina artificial é de fato inteligente. De acordo com o teste, se não conseguirmos identificar se estamos conversando com uma pessoa ou um computador, então podemos concluir que o computador apresenta uma inteligência genuína semelhante à inteligência humana. Searle pretende demonstrar que esse teste não é uma garantia para se identificar uma inteligência artificial que possa de fato compreender alguma coisa. Vamos supor que temos um programa escrito e processado com a função de simular o entendimento do idioma chinês,e que, em vista disso, esse programa responda questões formuladas em chinês através da exibição de respostas expressas em frases também em chinês. Ainda que as frases de resposta sejam tão adequadas quanto às de um falante chinês humano, não poderemos descobrir que esse programa realmente entende chinês apenas avaliando por comparação os desempenhos do programa e do falante humano, tal como prescreve o teste de Turing. De acordo com Searle, imaginemos um homem dentro de um quarto por onde entram seqüências escritas de caracteres chineses. Esse homem hipotético não possui nenhuma compreensão de chinês, mas, apesar disso, dentro do quarto há um livro escrito em inglês com instruções sobre como manipular as seqüências de caracteres chineses introduzidas no quarto. Seguindo as instruções do livro, o homem escreve e envia para fora do quarto outras seqüências de caracteres chineses. O livro de instruções estabelece regras para a manipulação dos símbolos chineses de modo absolutamente sintático. Essas regras poderiam ser estabelecidas logicamente, na forma “se, então”, como por exemplo: se for introduzido um ideograma chinês com determinado formato, então escreva e envie um ideograma com um outro formato específico. Embora o homem não saiba disso, os caracteres que entram são perguntas feitas por pessoas fora do quarto; já os caracteres enviados para fora são respostas às perguntas formuladas. Considerando que esse homem consiga habilmente manipular com rapidez os caracteres chineses, o que temos nessa situação é um conjunto de perguntas e respostas escritas em chinês entrando e saindo do quarto, gerando um padrão ou um programa que se assemelha à compreensão de um

idioma, mas, ao contrário, não ocorre nenhum entendimento verdadeiro. Searle conclui:

> Em virtude da realização de um programa formal de computador, do ponto de vista de um observador externo, alguém se comporta exatamente como se entendesse chinês, mas de qualquer modo não compreende uma só palavra de chinês. (SEARLE, 1984, p. 32).

Não apenas o programa de simulação de chinês não pode por si só assegurar uma compreensão de chinês, mas também nenhum tipo de computador digital, construído com qualquer tecnologia possível, pode compreender qualquer idioma ou linguagem. O entendimento de um sistema lógico de comunicação simbólica permanece para sempre além dos limites de um computador. Isso ocorre devido ao fato de que o computador não possui nenhuma característica especial que porventura o homem do quarto não apresente. Um computador possui apenas um programa abstrato para processar caracteres chineses que são fornecidos como dados de entrada. Um computador pode ser dotado de sintaxe, mas é desprovido de semântica. Ora, o entendimento de qualquer coisa no mundo depende de uma capacidade semântica. Para se compreender intencionalmente um idioma como por exemplo o chinês, é preciso uma capacidade suplementar além da simples sintaxe. Apresentar um estado mental, ainda que seja um estado de entendimento intencional de uma linguagem, não demanda somente um escopo de símbolos formalmente definidos. É preciso que se tenha um significado diretamente correlacionado a esses símbolos. Contudo, de acordo com a argumentação do quarto chinês, um dispositivo computacional tem somente símbolos formais porque suas atividades se definem completamente em termos de sua capacidade de memória e processamento de dados para a implementação de um programa previamente especificado pelos projetistas de *software*. E, um programa, apesar de ser capaz de realizar as mais variadas funções, é definido apenas de acordo com suas regras sintáticas, não apresentando nenhuma forma de semântica. Não é possível atribuir intencionalidade a um programa de computador se suas atividades são definidas em termos de processamento formal de informação.(SEARLE,1980).

O processamento não poderia, segundo essa concepção, reproduzir o conteúdo semântico dos estados mentais.

Pode-se comparar o fundamento do argumento do quarto chinês se compararmos duas situações teóricas diferentes entre si, a saber: se alguém for questionado sobre algum assunto em uma língua conhecida (no caso em questão, o inglês) e, em contrapartida, se esse mesmo alguém for questionado e produzir respostas em uma língua que seja desconhecida no tocante ao significado das sentenças proferidas. Podemos procurar estabelecer qual a diferença essencial dessas duas situações, se, de acordo com Searle, a uma pessoa específica que estiver dentro de um quarto for feito um interrogatório a respeito de pequenas trivialidades sobre a vida cotidiana dessa pessoa. As questões colocadas poderiam ser relacionadas a assuntos simples, tais como a cidade em que a pessoa reside, qual o tipo de filme a que a pessoa gosta de assistir, qual a sua atividade profissional, e assim por diante. Se as perguntas formuladas durante o interrogatório estiverem enunciadas em inglês, e, além disso, as respostas proferidas pela pessoa estiverem enunciadas também em inglês, a diferença entre essa situação e um interrogatório em que é apresentado um conjunto de perguntas e respostas em uma língua não compreendida (no caso, temos o exemplo das questões introduzidas no quarto em forma de caracteres chineses), reside justamente no fato de que as perguntas em inglês são entendidas porque estão sendo expressas, seja de maneira escrita ou simplesmente oralmente, em uma cadeia de símbolos cujos significados são conhecidos pela pessoa do exemplo em causa. Analogamente, quando a pessoa enuncia um conjunto de frases em inglês como resposta às perguntas formuladas no interrogatório, os símbolos constituintes das palavras da língua inglesa são corretamente compreendidos pois apresentam significado para a pessoa que fala o idioma inglês. Na situação imaginária do quarto chinês não ocorre um situação significativa, mas, ao contrário, o que está acontecendo é tão somente uma manipulação de símbolos formais semelhante ao processamento informacional realizado por um programa de computador, sendo que nenhum significado está sendo atribuído às seqüências de caracteres ou ideogramas chineses.

O argumento de Searle pode ser analisado através da compreensão de suas premissas lógicas:

1) programas são sintáticos;
2) mentes têm conteúdo semântico;
3) a sintaxe não é suficiente para a semântica;

A primeira premissa nada mais é do que uma definição de um programa de computador, a saber: um programa computacional é um conjunto definido de instruções de manipulação sintática previamente especificadas.

A segunda premissa é uma especificação da natureza de uma mente de acordo com a definição de Searle. Todo estado mental, enfim todo o conjunto de pensamentos, crenças e desejos, objetivos, exibe um fenômeno que filosoficamente é denominado de intencionalidade. O significado disso é que todo estado mental possui é de certa forma direcional, ou seja, um estado mental é sobre alguma coisa no mundo. Um estado mental, em síntese, se refere a algum estado de coisas específico. O motivo de um estado mental se referir a algum estado de coisas é devido ao próprio conteúdo intrínseco de um estado mental, conteúdo esse que dirige inevitavelmente o estado mental para alguma coisa no mundo.

A terceira premissa do argumento de Searle é uma proposição que procura estabelecer a distinção entre o que é puramente formal, e em contrapartida, o que possui conteúdo tal como um estado mental.

Se considerarmos as premissas do argumento, perceberemos que o foco central da argumentação depende de a terceira premissa ser verdadeira ou falsa. Apenas se aceitarmos a noção de que a sintaxe não é suficiente para a semântica, poderemos aceitar a conclusão de Searle de que um computador não pode apresentar uma intencionalidade apenas pelo fato de instanciar um programa correto. Searle não aceita que a execução de um programa sintático

possa resultar em intencionalidade porque acredita que o conteúdo de um estado mental esteja relacionado à supostas propriedades causais do cérebro humano. Contudo, no dizer de Dennett:

> [...] para Searle, o cérebro não é um computador; pois tem "propriedades causativas" do tipo que nenhum computador possui. Searle não fornece nenhuma explicação sobre o que são essas propriedades causais e tampouco como as ciências físicas poderiam compreender tais propriedades. (DENNETT, 1995, p. 445)

Searle conclui que a maneira como as conexões neurológicas do cérebro causam o conteúdo intencional não pode se basear em uma forma de programação sintática de computador. Os cérebros são sistemas biológicos equivalentes aos sistemas digestivo, circulatório, respiratório, e assim por diante. Searle considera que o conteúdo de um estado mental é um fenômeno biológico cerebral equivalente a qualquer outro processo fisiológico. Dessa maneira, uma vez que a estrutura fisiológica intrínseca do tecido cerebral é de suma importância no que diz respeito às propriedades causais responsáveis pela presença dos estados mentais, Searle reitera a concepção de que a operação de um programa de computador não pode originar uma mente inteligente dotada de intencionalidade. A intencionalidade é um fenômeno biológico intrínseco, e a definição de um programa desconsidera a noção de propriedades estruturais intrínsecas, visto que um programa é definido unicamente em termos de uma organização funcional em um sistema de processamento informacional. Searle considera que alguns sistemas físicos, como o cérebro, têm as propriedades causativas relevantes para o pensamento e a intencionalidade. Dessa forma, qualquer projeto de construção de um artefato capaz de pensar deve basear-se não na noção de programas de computador, mas sim em um projeto de pesquisa que procure elaborar um sistema com propriedades causais semelhantes às propriedades cerebrais.

Um aspecto da crítica elaborada por Searle consiste em que um programa de computador não pode em princípio apresentar uma verdadeira

intencionalidade porque a própria noção de um programa não é intrínseca a um conjunto de objetos. Trata-se de que, para definirmos algo como um programa, precisamos antes de tudo da interpretação de uma intencionalidade externa ao programa:

> A múltipla realizabilidade de processos computacionalmente equivalentes em meios físicos diferentes não é somente um sinal de que os processos são abstratos, mas também de que eles não são intrínsecos ao sistema; eles dependem de uma interpretação de fora. (SEARLE, 1992, p. 209).

Em tese, qualquer objeto pode ser interpretado de maneira computacional. De acordo com os princípios da Turing- computabilidade:

1) Para um dado objeto, existe uma definição do objeto em questão tal que, de acordo com essa definição o objeto é um computador;

2) Para qualquer programa ou algoritmo bem como para um objeto adequado, existe uma definição do objeto segundo a qual o objeto está executando qualquer algoritmo programável;

De acordo com a especificação de um programa de computador, podemos complementar os dois princípios enunciados acima com a conclusiva tese de Church:

3) Qualquer algoritmo programável ou função executável pode ser implementado por uma máquina de Turing.

Pode-se dizer, por exemplo, que um determinado bloco de concreto está rodando um editor de texto devido ao fato de que os átomos de sua estrutura

molecular estão em um estado equivalente a um estado de máquinas de Turing. Ou ainda podemos encarar uma cadeira como um programa de computador. Todavia, a intencionalidade (e, por extensão todos os estados mentais) não pode ser interpretada dessa maneira porque é uma capacidade irredutível do cérebro. Complementando a conclusão do argumento do quarto chinês, Searle alega que um programa não pode exibir intencionalidade porque suas atividades não conseguem reproduzir o aspecto absolutamente irredutível da mente. Um computador, ainda que possa executar um programa para decodificar informações e processá-las na forma de sinais elétricos em seus circuitos integrados, não dispõe de nenhuma semântica para interpretar a informação que está sendo computada.

As refutações do argumento do quarto chinês envolvem a princípio o problema da escala de tempo: ao supor que um homem pudesse simular manualmente dentro do quarto as instruções escritas em chinês, Searle está na verdade propondo uma situação praticamente irrealizável. Há uma enorme diferença de níveis de complexidades entre uma simulação feita por um homem e um programa computacional de entendimento de um idioma. Um ser humano não poderia executar um programa como o proposto na experiência do quarto chinês, pois não escreveria os caracteres com a rapidez necessária considerando o tempo disponível e, além disso, a própria manipulação de instruções seria uma dificuldade prática para a efetuação do programa. Uma simulação feita por um ser humano é lenta e simples se comparada a um programa de computador real, e Searle, através de seu argumento, está desacelerando o processo de manipulação simbólica até um ponto onde não há entendimento intencional, embora no mundo real a relação entre a manipulação de símbolos formais e a intencionalidade ou consciência é rápida e muito mais complicada. Um programa capaz de entender chinês precisaria analisar todos os ideogramas que formam o *input*. De acordo com Dennett:

> Um programa de computador que pudesse realizar o processamento lógico da gramática das palavras em chinês constituintes dos dados de entrada seria tão demasiado complexo que não se poderia dizer, *a priori*, que o programa não apresente consciência. (DENNETT, 1991, p. 436)

O argumento de Searle, deve-se salientar, faz um apelo direto às nossas intuições ao fazer com que um ser humano que executa instruções seja identificado com um programa sofisticado de IA. Ainda que essa identificação seja forçada, no sentido das diferenças práticas entre as ações de um homem e a execução de um programa, Searle pretende fazer com que pensemos que um programa não pode entender chinês apenas devido ao fato de que um homem capaz de agir intencionalmente não está compreendendo chinês ao cumprir as instruções do quarto.

Uma maneira de refutar o argumento do quarto chinês consiste em fazer uma comparação entre níveis do sistema (DENNETT, 1991). Isso significa que embora o homem dentro do quarto não entenda chinês, esse homem, apesar de ser ele mesmo uma unidade intencional, pode ser visto com sendo parte do sistema do quarto. Podemos comparar esse homem a uma unidade do disco rígido de um computador. O disco rígido é apenas parte do sistema computacional. Se considerarmos o sistema da experiência do quarto chinês, o que inclui o homem, o livro de instruções e as seqüências de caracteres que entram e saem do quarto, poderemos concluir que todo o sistema pode certamente compreender semanticamente o significado das instruções elaboradas em chinês.

Como corolário dessas observações, podemos dizer que as células neurológicas não entendem um idioma , apesar de que as pessoas, enquanto sistemas biológicos complexos formados por um complicado aparato molecular funcional, são capazes de compreender um dado idioma. Desse modo, seria um erro concluir que pessoas não podem entender um idioma somente pelo fato de que os componentes neurológicos não entendem um idioma lingüístico.

Ainda em relação à refutação do quarto chinês baseada em níveis de sistema, Dennett pretende mostrar o erro do argumento de Searle propondo a noção de níveis de implementação (HOFSTADTER e DENNETT, 1981). De acordo com essa noção, um sistema pode simular um outro sistema qualquer, assim como uma máquina universal de Turing pode simular qualquer outra máquina de Turing. Se tivermos um conjunto de máquinas sistêmicas

simuladas estruturalmente, veremos que a junção de uma máquina à outra resulta no que é denominado máquina virtual. Em uma estrutura hierárquica de máquinas virtuais, uma determinada máquina sempre é simulada consecutivamente por outra máquina em cada nível da hierarquia. Embora não ocorra entendimento em um dado nível isoladamente (note-se que células neurológicas, enquanto máquinas sistêmicas, não compreendem um idioma, embora uma pessoa, como outro nível de máquina, possa na prática compreender um idioma), poderia haver uma forma de interligação de informação entre os níveis de máquina. Essa conexão entre níveis de máquina seria a maneira como o programa do quarto chinês consegue alcançar a compreensão do idioma chinês. O entendimento intencional pertence ao quarto enquanto sistema, e não devemos imputar a compreensão a um homem que, sendo um nível sistêmico que não entende o chinês, apenas utiliza o livro de instruções em inglês para traduzir os ideogramas do quarto.

Uma vez que o argumento do quarto chinês se concentra apenas em um nível sistêmico (no caso, o homem), podemos a princípio pensar que um processo de computação não pode conduzir à intencionalidade. Contudo, uma vez que se considere que os níveis interferem entre si em uma cadeia de máquinas, pode-se concluir que é possível que um programa seja intencional. O ponto central é que, ao contrário do que alega Searle, não é necessária uma intencionalidade externa ao programa. A própria interação das máquinas em variados níveis de sistema, operando de acordo com as leis da física, pode instanciar um programa adequado para produzir artificialmente uma verdadeira forma de intencionalidade.

A experiência do quarto chinês consiste em uma simulação de compreensão de chinês. Pode-se fazer simulações computadorizadas de qualquer processo se for fornecido ao programa de computador um conjunto de informações escritas para descrever o processo em causa. É possível por exemplo, fazer simulações de análises econômicas em uma bolsa de valores. Também é possível fazer simulações de diagnósticos médicos, estratégias militares, resultados de eleições políticas, fenômenos naturais como tempestades e furacões, experiências científicas e aeronáuticas, e assim por diante. Searle pretende concluir que a IA Forte é impossível porque uma

simulação por computador de processos cerebrais não possuiria efetivamente nenhum estado mental, assim como não se supõe que uma simulação de tempestades apresentaria água líquida que possa molhar algum objeto.

Como uma forma de apresentar uma refutação desse raciocínio, Dennett primeiramente enfatiza (DENNETT,2006, p 262) que, embora autores como Searle tentem provar que os computadores não possam, em tese, reproduzir atividades e capacidades, como, por exemplo, o raciocínio humano, o programa de pesquisa de IA conseguiu, a despeito de argumentos céticos, obter êxito em atividades consideradas inteligentes, tais como o jogo do xadrez ou mesmo a prova de alguns teoremas matemáticos.O erro de Searle é supor que os pesquisadores de IA consideram que uma simulação de computador seria idêntica aos aspectos e estados de coisas no mundo real. Contudo, como diz Dennett:

> A estratégia de simulação computadorizada com freqüência tem sido mal compreendida. Na simulação computadorizada nunca é o caso de um modelo de computador ser *indistinguível* daquilo que está sendo modelado. (DENNETT,2006, p 262)

Não se espera que uma simulação computadorizada de um tornado possa arrastar algum objeto. Supor que isso possa acontecer é, de acordo com Dennett (DENNETT,2006, p 262) “ semelhante a um erro entre uso e menção, com se agachar diante da palavra “leão””. Um programa de simulação de um tornado, por exemplo, consiste em um programa que fornece descrições do comportamento de um tornado quando lhe são fornecidos dados das condições meteorológicas próprias de um tornado. As descrições que formam o *input* e o *output* poderiam ser em qualquer linguagem de programação, desde que incluíssem dados sobre nuvens, ondas marítimas, pressão atmosférica, e qualquer outro dado científico sobre o estado dos tornados.

Um programa que simulasse um tornado, deve, em síntese, fornecer boas predições sobre os efeitos físicos de um tornado em uma ampla variedade de condições ambientais altamente complicadas. Uma simulação de computador é, dessa forma, equivalente a uma série de condicionais do tipo *se*

as condições A,B,C ocorrerem, então haverá as conseqüências D,E,F. Podemos resumir esses condicionais na forma de conectivos lógicos como *se P, então Q*. A maneira apropriada pela qual as condicionais de um programa de simulação podem ser coligidas seria pela apreensão das variáveis contingentes derivadas das leis gerais do fenômeno a ser simulado. Se um programa de simulação de tornados for escrito, deve-se primeiramente ter uma teoria geral sobre o comportamento dos tornados, sendo que essa teoria deve ser satisfatória no sentido de prover informações relevantes sobre as condições meteorológicas prováveis que acarretam a formação de um tornado. A estratégia de simulação computadorizada é relevante porque nos permite avaliar os aspectos incorretos de uma teoria, bem como procurar entender alguns dados que porventura não tenham sido considerados. Uma vez que se tenha um programa escrito e processado, as implicações de seu funcionamento podem ser determinadas e estudadas pelos pesquisadores. Podemos considerar que um programa é um tipo de experimentação para avaliarmos nossas hipóteses sobre um fenômeno qualquer. Podemos, por exemplo, obter uma apreensão rápida e em tempo real sobre o impacto de um tornado em uma localidade onde existem muitas residências construídas. Um programa de simulação por computador consiste, resumidamente, em um gerador de conseqüências que uma teoria atribui a variadas condições. Ocorre que, em uma simulação, é importante salientar a diferença entre a teoria que dirige o programa e o modo de apresentação do conjunto de dados de entrada/saída. Pode-se, por exemplo, apresentar os dados correspondentes ao formato de um tornado como sendo simplesmente um feixe luminoso em forma de cone em uma tela de vídeo. O feixe não é idêntico ao formato de redemoinho de um tornado. O que importa em uma simulação é se a teoria que estamos apreendendo é adequada ou não para o estudo do fenômeno que estamos modelando pelo computador, A teoria corporificada no programa na forma de dados dirige os efeitos da apresentação de uma simulação, sejam esses efeitos apresentados na forma de uma imagem de vídeo ou qualquer outra forma. Mas a única função interpretável da apresentação simulada de um fenômeno é tão somente mostrar adequadamente a simulação para o escrutínio dos pesquisadores. A apresentação simulada não é de forma alguma uma reprodução do mundo real.

A relevância dessa discussão para a questão sobre a IA é que uma forma de inteligência artificial não consistiria apenas em uma simulação, mas em uma instanciação que possa reproduzir todos os aspectos importantes do mental e apresentar intencionalidade. Desse modo, embora Searle esteja certo ao dizer que uma mera simulação de eventos mentais não seria suficiente para a produção de um artefato dotado de intencionalidade, é possível que um programa de computador possa apresentar atividade intencional instanciando ou reproduzindo atividades análogas às do cérebro humano.

CONCLUSÃO

A noção de intencionalidade é um tema de profundo interesse para a investigação filosófica. Desde as primeiras investigações realizadas pela corrente fenomenológica, o conceito de intencionalidade, enquanto propriedade dos estados mentais, tem sido esclarecido através de diferentes abordagens apresentadas pela filosofia contemporânea.

A filosofia da mente, enquanto correlacionada à corrente da filosofia analítica, propôs ao longo das últimas décadas algumas soluções teóricas visando a compreensão da intencionalidade. Algumas dessas soluções foram mencionadas ao longo da presente dissertação. Temos, primeiramente, a teoria da identidade. Como uma forma de fisicalismo, a teoria da identidade procurava solucionar o problema da intencionalidade tão somente como uma identificação dos estados mentais com os estados neurológicos do cérebro. Em contrapartida, o funcionalismo de máquinas de Turing não fazia nenhuma referência aos estados físicos do cérebro: a intencionalidade consistiria tão somente em um estado lógico abstrato de um sistema de computador.

Mais recentemente, foram sugeridas outras soluções para o entendimento da intencionalidade. Entre essas teorizações, uma dentre as mais conhecidas é a abordagem conexionista e as assim chamadas redes neurais artificiais (CHUCHLAND;SEJNOWISKI, 1992). Essas redes seriam capazes de reproduzir os estados mentais na medida em que são estruturadas de maneira análoga ao cérebro humano, ou seja, através da ligação em forma de rede entre as unidades de processamento, como os neurônios.

O pensamento de Dennett, diferentemente das abordagens mencionadas, é baseado em uma forma de funcionalismo que, embora mantenha uma neutralidade em relação ao substrato físico de uma entidade, está estritamente ligado á noção de uma postura ou estratégia de predição. Para Dennett, o funcionalismo pode ser salvaguardado não através da identificação dos termos intencionais com os estados lógicos de máquina de

Turing, mas sim através da adoção, em relação a um dado sistema, de uma estratégia de predição:

> "[...] decide-se tratar o objeto cujo comportamento se quer prever como um agente racional; depois, imaginam-se que crenças esse objeto deve possuir, dado a sua posição no mundo, bem como seus objetivos. Imaginam-se também os desejos que deveriam motivar o objeto,com base nas mesmas considerações, e finalmente, prevê-se que esse agente racional agirá de determinada maneira, visando realizar seus propósitos, de acordo com suas crenças. Um pouco de raciocínio a partir do conjunto escolhido de crenças e desejos resultará em muitas, embora não em todas, ocasiões uma decisão sobre o que o agente deveria fazer; essa decisão é o que conseguimos prever que o agente irá fazer" (DENNETT,1987, p.17)

O vocabulário intencional, nesta acepção, é apenas uma ficção útil, assim como o conceito de "centro de gravidade". Os físicos, em seu próprio domínio, obtêm êxito explanatório postulando um hipotético ponto onde podemos supor que se localiza o peso de um objeto. Por analogia, a intencionalidade é uma construção que atribuímos aos entes com o objetivo de prever seu comportamento. Na medida em que conseguimos alcançar êxito na predição das ações de um objeto através da atribuição a esse mesmo objeto de termos como crenças e desejos, podemos dizer que é útil considerar tal objeto como um ser racional, pois seu desempenho correspondeu adequadamente às nossas pressuposições.

Dennett considera que uma compreensão científica da intencionalidade seja totalmente possível em princípio. Todavia, podemos afirmar à guisa de conclusão que mesmo uma elucidação sobre todos os aspectos empíricos do comportamento humano não seria uma justificativa para adotarmos a postura física, em lugar da postura intencional, na tentativa de predizer as ações humanas. Para sistemas complexos como o cérebro humano, a postura intencional é a estratégia mais prática possível para prevermos o comportamento.

Essas considerações se aplicam também aos sistemas de computador desenvolvidos pela pesquisa em inteligência artificial. Esses sistemas alcançaram um grande nível de complexidade e sofisticação no desempenho de tarefas como jogar xadrez ou realizar análises econômicas. Assim sendo, a melhor forma de predizer satisfatoriamente o comportamento desses sistemas é através da atribuição, em relação a essas máquinas, da postura intencional, considerando esses sistemas como possuidores de intencionalidade.

Essa perspectiva foi criticada através de argumentos tais como o quarto chinês e o teorema de Gödel. Esses argumentos consideram que a hipótese de computadores apresentarem comportamento intencional é um erro filosófico. Conforme foi argumentado no terceiro capítulo da dissertação, essas críticas são ambas falaciosas. A partir do momento que for projetado um sistema computacional que execute um programa funcionalmente apropriado, pode-se dizer que esse programa possa apresentar intencionalidade.

Em suma, o conceito de postura intencional, proposto por Dennett, é um tipo de abordagem materialista da intencionalidade. De fato, a estratégia da postura intencional é coerente com uma perspectiva mecanicista. Uma vez que se trata de uma abordagem naturalista da intencionalidade, pode-se concluir que sistemas de inteligência artificial podem em princípio apresentar um comportamento intencional análogo ao comportamento humano. O aspecto abstrato da postura intencional faz com que essa estratégia possa ser aplicada tanto às máquinas quanto aos seres humanos.

BIBLIOGRAFIA

ABBAGNANO, N. *Dicionário de Filosofia*. São Paulo : Martins Fontes, 2007.

ARMSTRONG,D. *A Materialist Theory of Mind*. New York: Routledge, 1968.

BLACKBURN, S*. Dicionário Oxford de Filosofia*. Rio de Janeiro : Jorge Zahar, 1997.

BODEN, M. *Atificial Intelligence and Natural Man*. Cambridge : MIT Press, 1987.

BRENTANO, F. *Psychology from an Empirical Standpoint* (1º ed. 1874). New York.: Routledge, 1995..

CHOMSKY, N. *Aspectos da Teoria da Sintaxe*. São Paulo : Abril Cultural, 1979.

_____ *Novos Horizontes no Estudo da Linguagem e da Mente*. São Paulo : UNESP, 2006.

CHURCHLAND, P. M. *Scientific Realism and Plasticity of Mind*. Cambridge : Cambridge University Press, 1979.

_____ *A Neurocomputational Perspective*. Cambridge: MIT Press,1992.

_____. *The Engine of reason: The Seat of the Soul*. Cambridge : MIT Press, 1995.

_____. *Matéria e Consciência*. São Paulo : UNESP, 2004.

CHURCHLAND, P. S. *Neurophilosophy*. Cambridge : MIT Press, 1986.

CHURCHLAND, P. S. , SEJNOWISKI, T. *The Computational Brain*. Cambridge : MIT Press, 1992.

DARWIN. C. *A Origem das Espécies*. São Paulo : Hemus, 1999.

DENNETT, D*. Brainstorms: Ensaios filosóficos sobre a mente a psicologia.* São Paulo UNESP,2006.

_____ *Consciousness Explained*. Boston : Little Brown, 1991.

_____ . *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York: Simon and Schuster,1995.

_____ *Tipos de Mentes: Rumo a um entendimento da consciência*. Editora Rocco, 1996.

_____ *The Intentional Stance*. Cambridge: MIT Press, 1987..

DESCARTES, R. *Discurso do Método*. São Paulo : Abril Cultural, 2001.

_____ *Meditações Metafísicas*. São Paulo : Abril Cultural, 2001.

EDELMAN, G. *The Remembered Present: A Biological Theory of Consciousness*. New York : Basic Books, 1989.

_____ *Bright Air, Brilliant Fire: On the matter of the Mind*. New York : Basic Books, 1993.

FLANAGAN, O. *The Science of the Mind*. Cambridge : MIT Press, 1991.

_____. *Consciousness Reconsidered*. Cambridge : MIT Press, 1992.

GÖDEL, K. *On Formally Undecidable Propositions of Principia Mathematica and related Systems*. tr. B. Meltzer. New York: Basic Books, 1962.

HOFSTADTER, D. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books, 1979.

HOFSTASTER, D DENNETT,D *The Mind's I: Fantasies and Reflections on Self & Soul*. New York: Basic Books, 1981.

HORGAN, J. *O Fim da Ciência*. São Paulo : Companhia das Letras, 1997.

HUISMAN, D. *Dicionário de Obras Filosóficas*. São Paulo: Martins Fontes, 2000.

_____ *Dicionário dos Filósofos*. São Paulo : Martins Fontes, 2001.

HUME, D. *Investigação sobre o Entendimento Humano*. São Paulo : Abril Cultural, 2001.

JAYNES, J. *Origin of Consciousness in the Breakdown of the Bicameral Mind.* New York : Mariner Books, 1976.

JOHNSON-LAIRD, P. *The Computer and the Mind*. Cambridge : Havard University Press, 1989.

LUCAS, J. "Minds, Machines and Gödel" In: *Philosophy* XXXVI, 1961, p. 112-127.

MCGINN, C*. The Problem of Consciousness*. Cambridge : Blackwell Publishing Ltd, 1993.

MINSKY, M. *A Sociedade da Mente*. Rio de Janeiro : Francisco Alves, 1992.

NAGEL, E ,NEWMAN, J. *Gödel's Proof*. New York. New York University Press, 1958.

PENROSE, R. *Emperor's New Mind Computers, Minds and the Laws of Physics.* Oxford : Oxford University Press, 1989.

_____ *Shadows of the Mind*. New York: Oxford University Press, 1994.

PUTNAM, H. *Mind, Language and Reality*. Cambridge : Cambridge University Press, 1975.

QUINE, W. V. *Ontological Relativity & Other Essays*. New York: Columbia University Press, 1977.

_____*Word and Object*. Cambridge : MIT Press, 1960.

RICH, E./KNIGHT, K. *Inteligência Artificial*. São Paulo : Makron Books, 1997.

RYLE, G. *The Concept of Mind*. London: Hutchinson, 1949.

SEARLE, J. *Intencionalidade*. São Paulo : Martins Fontes, 1995.

_____"Minds, Brains and Programs" In: *Behavioral and Brain Sciences* 3, 1980 , 417-424.

_____ *Minds, Brains and Science*. Cambridge: Harvard University Press, 1984.

_____.. *O Mistério da Consciência* Rio de janeiro : Paz e Terra, 1998.

_____. *Rediscovery of Mind*. Cambridge: MIT Press, 1992.

SIMON, H. *The Sciences of the Artificial*. Cambridge : MIT Press, 1969.

SKINNER, B. F. *About behaviorism*. New York: Vintage Books, 1974.

_____ *Contingências do Reforço*. São Paulo : Abril Cultural, 1979.

_____ *Para Além da Liberdade e da Dignidade*. Lisboa : Edições 70, 2000.

_____ *Ciência e Comportamento Humano*. São Paulo: Martins Fontes, 2003.

TREFIL, J. *Somos Diferentes?* Rio de Janeiro : Rocco, 1999.

TURING, A. " Computing Machinery and Intelligence" In: *Mind* LIX (236) 433-460.

WANG, H. *Logical Journal: From Gödel to Philosophy*. Bradford Book, 1997

WIENER, N. *Cybernetics or Control and Communication in the Animal and the Machine.* Cambridge : MIT Press, 1965.