Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação

Análise da Arquitetura Baars-Franklin de Consciência Artificial Aplicada a uma Criatura Virtual

Autor: Ricardo Capitanio Martins da Silva
Orientador: Prof. Dr. Ricardo Ribeiro Gudwin

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: **Automação**.

Banca Examinadora

Fernando Antônio Campos Gomide, Dr. ..... DCA/FEEC/UNICAMP
Fernando Von Zuben, Dr. ..... DCA/FEEC/UNICAMP
Henrique Elias Borges, Dr. ..... DECOM/CESET-MG
João Eduardo Kogler Junior, Dr. ..... POLI/USP
Ricardo Ribeiro Gudwin, Dr. ..... DCA/FEEC/UNICAMP

Campinas, SP
Julho/2009

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP
(Será construída após a defesa)

Silva, Ricardo Capitanio Martins da
F149u Sobre a Preparação de Teses e Dissertações
na Faculdade de Engenharia Elétrica e de Computação
Ronny Guimarães Márs ico. – Campinas, SP:
[s.n.], 2003.

Orientador: Ricardo Ribeiro Gudwin.
Tese (mestrado) - Universidade Estadual de Campinas,
Faculdade de Engenharia Elétrica e de Computação.

1. Keyword1. 2. Keyword2. 3. Keword3. 4. Arquitetura de
sistemas (Computação). 5. Agentes inteligentes (Software).
6. Sistemas operacionais distribuídos (Computadores).
I. Gudwin, Ricardo Ribeiro. II. Universidade Estadual de Campinas.
Faculdade de Engenharia Elétrica e de Computação. III.
Título

Resumo

Há muito tempo, a consciência humana tem desafiado cientistas de várias áreas do conhecimento. Em computação, na última década, tem se observado um crescimento significativo dos estudos sobre consciência artificial. Uma abordagem computacional promissora é a da arquitetura Baars-Franklin, desenvolvida por Stan Franklin. Inspirada na teoria do *workspace* global de Bernard Baars, essa arquitetura foi aplicada na criação de agentes autônomos, como um tutor virtual para auxiliar o treinamento de astronautas no uso um braço mecânico da ISS e um agente distribuidor de trabalhos aos marinheiros dos EUA. Esse trabalho visa examinar essa tecnologia e verificar as vantagens e desvantagens do seu uso na implementação de agentes inteligentes. Como o estudo de consciência artificial é inerentemente multidisciplinar, inicialmente apresenta-se as principais teorias gerais de consciência. Em seguida, é feito um levantamento dos principais trabalhos na área de consciência artificial. Por fim, utiliza-se a arquitetura Baars-Franklin no controle de uma criatura virtual em um problema de navegação autônoma. Essa dissertação traz uma série de contribuições teóricas, agrupando teorias de outras áreas do conhecimento fundamentais para o desenvolvimento de agentes com consciência artificial, clarifica a arquitetura Baars-Franklin e realiza um estudo de caso prático da aplicação desse modelo na construção de um agente autônomo.

Palavras-chave: consciência artificial, sistemas cognitivos, cognição artificial.

Abstract

For years, the human consciousness have challenged scientists from several areas of knowledge. In computer science, in the last decade, it has been observed a significant growth in the number of studies about artificial consciousness. The architecture Baars-Franklin, developed by Stan Franklin, is a computational promising approach. Inspired in the Bernard Baars' global workspace theory, this aarchitecture was applied in the development of autonomous agents, such as a virtual tutor to support of astronauts training on the manipulation of the ISS robotic arm and an agent to assign billets to USA Navy sailors. This work aims to discuss this technology and verify its advantages and disadvantages. Due to the intrinsic multidisciplinarity characteristic of the study of artificial consciousness, first of all, it is presented the main general theories about consciousness. After that, it is surveyed the principal studies about artificial consciousness. Finally, the architecture Baars-Franklin is applied to control a virtual creature in an autonomous navigation problem. This dissertation brings some theoretical contributions, clustering theories from other areas of knowledge, clarifies the architecture Baars-Franklin and shows a practical case study of the application of this model in order to build an autonomous agent.

Keywords: artificial consciousness, cognitive systems, artificial cognition.

À minha mãe Terezinha e
ao meu pai Oscar.

Morpheus: We have only bits and pieces of information, but what we know for certain is that some point in the early twenty-first century all of mankind was united in celebration. We marvelled at our own magnificence as *we gave birth to AI*

Neo: AI - you mean Artificial Intelligence?

Morpheus: A *singular consciousness* that spawned an entire race of machines

from The Matrix, 1999

Agradecimentos

Sonho que se sonha só...
é só um sonho que se sonha só...
Mas sonho que se sonha junto é
realidade.

Raul Seixas

A todos que me auxiliaram na construção desse trabalho.

Ao CNPQ, pelo apoio financeiro.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

> If you were designing an organic machine to pump blood you might come up with something like a heart, but if you were designing a machine to produce consciousness, who would think of a hundred billion neurons?
>
> Jonh R. Searle

1.1 Prólogo

O entendimento da consciência, um fenômeno aparentemente inerente à vida mental dos seres humanos, tem se demonstrado um enigma intricado a físicos, psicólogos, matemáticos, filósofos, médicos e neurocientistas. Mesmo com todo a compreensão existente sobre o funcionamento físicoquímico do cérebro, a consciência carrega algo misterioso quando se tenta responder à pergunta de como a consciência é produzida pelos processos do cérebro (Wilson & Keil, 1999, p. 191). Os estudos de consciência, usualmente multi-disciplinares, buscam entender a natureza da consciência, suas propriedades, sua função, e as vantagens para os seres que a possuem. A área de pesquisa está em um estado pré-paradigmático e ainda não há um consenso sobre os métodos, modelos e protocolos para se realizar pesquisas sobre consciência.

Há poucos mais de uma década, tem se observado um aumento crescente nas pesquisas que buscam agrupar consciência e computação, principalmente através de testes de teorias de consciência através do uso de modelos computacionais (Franklin *et al.* , 1998; Dehaene *et al.* , 2003; Ramamurthy *et al.* , 2006; Dubois, 2007a; Gamez, 2008b; Shanahan & Connor, 2008). Também especula-se que esses estudos poderiam levar a criação de máquinas mais inteligentes, ou por terem comportamentos mais próximos aos dos seres humanos, ou por possuirem algoritmos de controle mais eficazes. Essa linha de pesquisa tem se tornado conhecida por *consciência artificial*[1] e é também chamada

[1] em inglês, *artificial consciousness*.

de *consciência de máquina*[2].

Consciência artificial, assim como acontece com os estudos gerais de consciência, é atualmente uma área de pesquisa bastante heterogênea. Alguns pesquisadores tem interesse em modelar características cognitivas da consciência, outros trabalham com comportamentos associados à consciência. Existem também aqueles que são interessados em modelar características cognitivas da consciência e ainda aqueles que procuram modelar a consciência fenomenológica[3] (Gamez, 2008a).

Dentre as diversas abordagens encontradas na literatura, a arquitetura computacional Baars-Franklin (ABF) tem se mostrado promissora diante de uma série de agentes que a utilizam. ABF, baseada na teoria do *workspace* global (Baars, 1988, 1997), foi implementada computacionalmente por Stan Franklin, da Universidade de Memphis (EUA). O primeiro agente implementado, utilizando a ABF, foi *CMattie* (Franklin & Graesser, 1999) que era capaz de fazer reserva de salas para eventos periódicos trocando e-mails com seus interlocutores. *IDA* (Franklin, 2003, 2005; Baars & Franklin, 2007), baseada em uma segunda versão da ABF, era um agente com a capacidade de preparar a escala de marinheiros e fazer a negociação das atribuições seguindo as políticas da marinha americana. Mais recentemente, Daniel Dubois da Universidade de Quebec (Canadá), utilizou a ABF no desenvolvimento de um sistema tutor consciente chamado *CTS - Conscious Tutoring System* (Dubois, 2007a), um agente para auxiliar o treinamento de astronautas no uso de um braço mecânico da Estação Espacial Internacional.

1.2 Motivação

Diante dos resultados apresentados pelos agentes que utilizam a ABF e de um cenário de crescimento na importância dos estudos de consciência artificial internacionalmente[4], esse trabalho se propõe a analisar a abordagem utilizada na ABF para verificar os pontos positivos e negativos da adição dessa tecnologia à construção de criaturas artificiais.

[2] em inglês, *machine consciousness*.

[3] Esses pesquisadores acreditam que é possível gerar estados conscientes em máquinas, como aqueles presentes nos humanos.

[4] No Brasil, os únicos trabalhos encontrados que tentam associar consciência e engenharia são apresentados em (do Valle Filho, 2003) e (Halfpap, 2005), da Universidade Federal de Santa Catarina. Ambas, teses de doutorado, são baseadas na teoria de consciência de Damário (ver seção 2.3.2) e procuram criar um modelo de consciência aplicável a robôs. do Valle Filho (2003) apresenta-se como um trabalho especulativo, sem implementações reais, enquanto que Halfpap (2005) até mostra uma aplicação de tratamento do balanço energético do robô, mas não apresenta algoritmos e resultados de testes na implementação. Além disso, os dois trabalhos não trazem uma contextualização da área de consciência artificial e não citam outros trabalhos de agentes conscientes, que já existiam na época dessas publicações.

1.3 Objetivos

Os objetivos desse trabalho são:

- contextualizar e estudar as principais teorias gerais de consciência e os modelos de consciência artificial, de forma a entender as diversas abordagens e suas implicações ao estudo de consciência artificial;
- verificar o estado da arte dos trabalhos de consciência artificial e levantar os seus principais benefícios ao estudo de agentes inteligentes;
- realizar experimentos de prova de conceito com um agente autônomo, utilizando a arquitetura Baars-Franklin, para averiguar, na prática, as vantagens e desvantagens dessa abordagem.

1.4 Estrutura da dissertação

Além desse capítulo introdutório, o capítulo 2, aprofunda a compreensão dos diversos significados do termo "consciência", discute os principais modelos gerais de consciência e os principais problemas filosóficos que envolvem o estudo da mente e da consciência. Esses tópicos são importantes principalmente para os leitores provenientes das ciências exatas, ou que não estão habituados com o caráter multidisciplinar dessa linha de pesquisa.

O capítulo 3, complementa o estudo teórico realizado no capítulo anterior, com a apresentação dos principais modelos de consciência artificial, com foco nas arquiteturas cognitivas. Também são apresentados alguns estudos de aplicações específicas.

No capítulo 4, detalha-se a arquitetura Baars-Franklin e assim se encerra o estudo teórico de modelos computacionais de consciência, iniciado no capítulo 2. Nesse capítulo são apresentados os principais componentes dessa arquitetura.

No capítulo 5, apresenta-se o problema exemplo utilizado na realização dos experimentos para análise da arquitetura. Esse capítulo apresenta os detalhes da implementação, as implicações e resultados da utilização da abordagem da arquitetura Baars-Franklin em uma criatura artificial.

As conclusões dessa dissertação e os trabalhos futuros são apresentados no capítulo 6.

Um mapa da dissertação pode ser visto na figura 1.1.

1.4.1 Como ler essa dissertação

Essa dissertação tem um caráter interdisciplinar, combinando conceitos de outras áreas do conhecimento e da engenharia. Alguns capítulos podem ser lidos isoladamente se o leitor tiver interesse nos tópicos nele apresentados. Em outro caso, alguns capítulos podem ser preteridos, caso o leitor já domine os conhecimentos expostos. Cada capítulo apresenta uma breve introdução para facilitar a leitura.

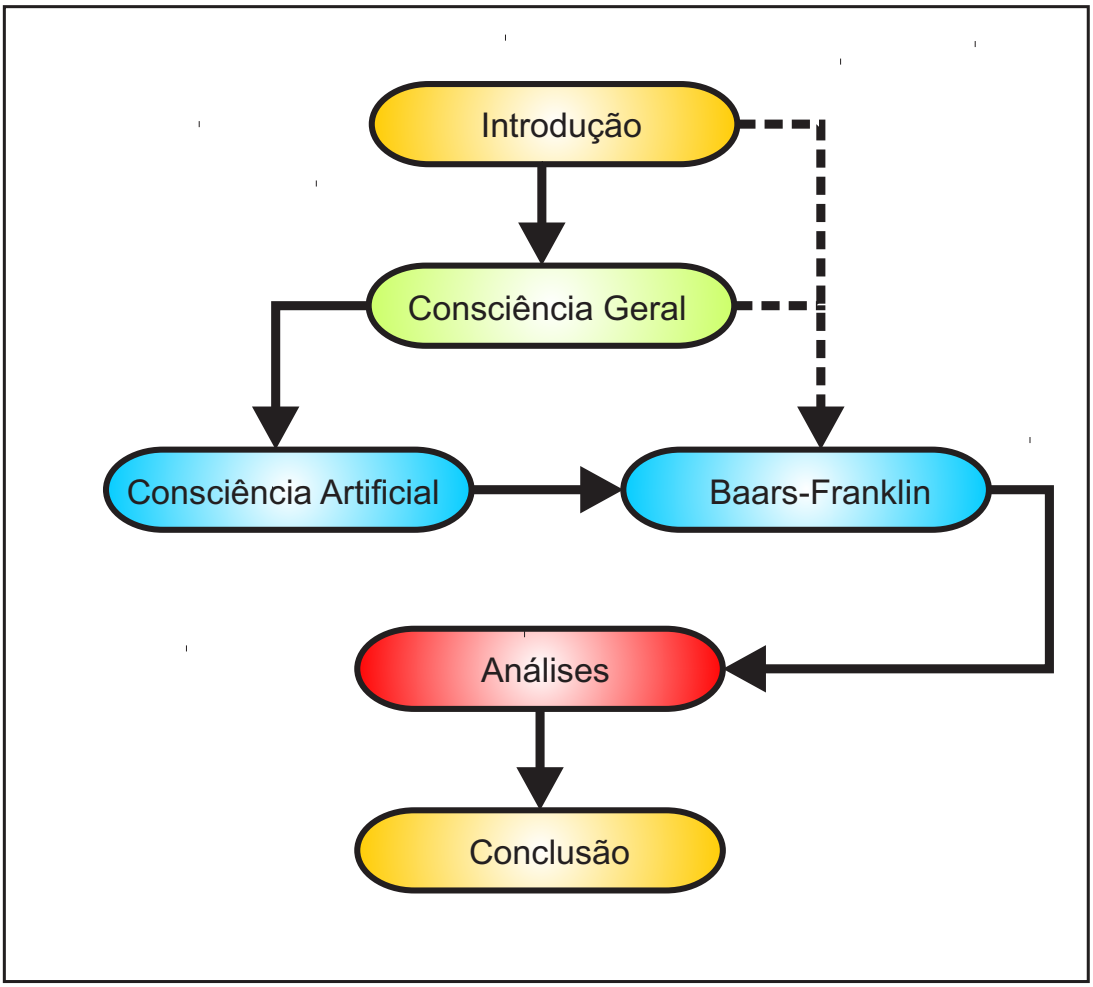


Figura 1.1: Mapa da dissertação. As setas tracejadas indicam caminhos alternativos de leitura.

Capítulo 2

Teorias Gerais de Consciência

> How it is that anything so remarkable as a state of consciousness comes about as a result of initiating nerve tissue, is just as unaccountable as the appearance of Djin, where Aladdin rubbed his lamp in the story...
>
> Julian Huxley

2.1 Introdução

O estudo da consciência tem caráter multidisciplinar e, por isso, é necessária uma revisão da pesquisa sobre o tema em diversas áreas do conhecimento. Esse capítulo, além de trazer diversas propostas envolvendo consciência, mostra de forma não exaustiva os principais modelos gerais de consciência encontrados na literatura.

Na seção 2.2, são discutidos os diversos significados da palavra *"consciência"*, abordando as variantes de Ned Block e alguns problemas relevantes da filosofia, como o *problema mente-corpo*, o *problema difícil da consciência* de Chalmers e a questão dos *qualia*. A seção 2.3, apresenta os principais modelos de consciência da atualidade, nas suas mais diversas vertentes. Muitos modelos computacionais se inspiram em características dos modelos gerais, daí a importância de uma revisão dessa literatura. Por fim, mesmo diante de diversos significados e variantes do que se entende por consciência, na seção 2.4, mostra-se algumas semelhanças e uma certa convergência entre os modelos de consciência estudados.

No capítulo 3, o estudo dos modelos de consciência será complementado com os chamados modelos computacionais da consciência, abordando os tópicos específicos de consciência artificial.

2.2 O que é Consciência?

Atualmente tem havido um crescimento significativo nos estudos sobre consciência nas mais diversas áreas do conhecimento: Filosofia, Física e Matemática, Psicologia, Neurobiologia, Ciências Cognitivas, Medicina, Computação, entre outras. Apesar dos muitos estudos terem pontos em comum, o que se nota é uma grande dificuldade na conceituação do que é o fenômeno da consciência humana e da sua natureza. Há até mesmo dúvidas sobre a sua real existência.

Conceitualmente, o termo *"consciência"* abrange uma vasta gama de significados, gerando entendimentos diferentes e, muitas vezes, pontos de vistas divergentes. Segundo Miller, "dependendo da figura de discurso escolhida ela é um estado de ser, uma substância, um processo, um lugar, um epifenômeno, um aspecto emergente da matéria, ou somente uma realidade verdadeira" (Miller, 1962, p.25). Ainda nessa linha, para Güzeldere, a consciência deve ser tratada como um conceito composto: "há simplesmente tantas conotações sob o termo que parece ser inútil tentar especificar um único conceito que cobriria todos os aspectos da consciência" (Güzeldere, 1997, p.22)[1].

Muitos ainda discutem a natureza da consciência, inclusive questionando se esta é realidade ou algo apenas aparente. Todas essas dúvidas e variantes levam a uma vasta diversificação no nível de descrição, de análise do problema, de métodos de pesquisa e protocolos de investigação, como será mostrado na seção 2.3.

2.2.1 As quatro consciências de Block

Seguindo a linha dos cientistas que acreditam que a *consciência* encapsula vários conceitos, Block (1995, 2002) defende a ideia dos quatro "tipos" funcionais: a *consciência fenomenal* que é aquela que carrega a inefável característica qualitativa do fenômeno consciente; *consciência de acesso* a qual é o fenômeno que temporalmente conecta recursos inconscientes, possibilitando que eles interajam entre si; *consciência de monitoramento* que está ligada aos processos que monitoram os sentidos dos estados internos; e a *autoconsciência* que se refere ao poder de separar o próprio indivíduo do ambiente.

Esses conceitos são interligados e nem sempre é possível distinguir claramente um do outro. Apesar disso, a classificação de Block delimita e formaliza a consciência de um ponto de vista computacional (Moura, 2006). A seguir cada tipo será analisado com mais detalhes.

Consciência Fenomenal (P-Consciência)

"Consciência fenomenal é experiência" já dizia Block (2002). A questão central da *P-Consciência* está na experiência qualitativa do fenômeno da consciência, o que faz o estado fenomenalmente consciente (*P-consciente*) possuir o conteúdo de "o que é experienciar" (Nagel, 1974) esse estado[2].

[1] Güzeldere descreve uma interessante revisão histórica e conceitual dos campos da filosofia e psicologia relacionados com consciência.

[2] A proposta de Nagel é retomada na seção 2.2.3.

Block deixa claro a dificuldade de dar uma definição não circular para a *P-Consciência* e acredita que isso não impeça o seu entendimento, apesar de, em alguns casos, dificultar a distinção entre os outros tipos funcionais. Assim, ele passa a citar alguns exemplos:

> Temos estados P-conscientes quando vemos, ouvimos, cheiramos, degustamos e temos dor. As propriedades P-conscientes incluem as características da *experiência* das sensações, sentimentos, e percepções e eu também incluiria pensamentos, desejos e emoções (Block, 2002).

Assim, a *P-Consciência* vai além do lado objetivo da percepção e da fisiologia pura e se relaciona aos *qualia* (veja seção 2.2.3). Por exemplo, ao beber uma taça de vinho, o estado *p-consciente* traria as sensações qualitativas da experiência, aquilo que o degustador tenta (muitas vezes com dificuldade) explicar a um terceiro ao provar o vinho. Nesse exemplo, é possível encontrar diversas teorias sobre o funcionamento do paladar, entretanto o lado subjetivo do problema ainda é uma questão em aberto.

Consciência de Acesso (A-Consciência)

A *A-Consciência* está ligada ao acesso à representação. Block (2002) diz que "uma representação é A-Consciente se ela for amplamente propagada[3] para o uso livre do raciocínio e para o controle 'racional' direto das ações (incluindo relatos)".

A *A-Consciência* está relacionada aos sensores, às crenças, aos estados e à representação, e, ainda, à forma como esses sensores refletem percepções do ambiente ou de si próprio. Mas possuir essas percepções não é suficiente, é preciso que as representações estejam disponíveis para um processamento de alto nível, como julgamento, raciocínio lógico, planejamento e controle racional das ações. Um exemplo de *A-Consciência* seria o caso em que, ao cortar o dedo, uma pessoa buscasse estancar o sangue e reduzir a dor, e seria capaz de dizer à outra o que fez e pedir por um curativo.

Autoconsciência (S-Consciência)

A *Autoconsciência* é a capacidade de ter a noção de si mesmo como algo distinto do resto do mundo (*self*) e ter a habilidade de utilizar esse conceito para pensar sobre si. Essa habilidade permite o autorreconhecimento (em frente a um espelho, por exemplo) e o conhecimento da próprio identidade, diferenciado-a do restante do ambiente.

A habilidade de se reconhecer em frente a um espelho é percebida em vários primatas, enquanto que nos cachorros, em um primeiro momento não, pois eles vêem as imagens refletidas como estranhos, acostumando-se com o passar do tempo. Block (2002) menciona um teste, que poderia ser considerado um teste de *autoconsciência* em que um ponto vermelho é fixado na testa de macacos. Os macacos entre 7 e 15 anos, ao passarem a frente do espelho, tentam retirar o ponto vermelho. Nos seres humanos isso só é observado ao final do segundo ano de vida.

[3] Nessa caso, a *"propagação"* pode ser entendida no sentido da teoria do *workspace* global de (Baars, 1988, 1997) em que há o *broadcast* de representações para todos os participantes, como será visto na seção 2.3.8.

Consciência de Monitoramento (M-Consciência)

A *M-Consciência* é a ideia da consciência como um tipo de monitoramento interno. Esse monitoramento pode tomar várias formas, apresentando-se como: *percepções internas*, como por exemplo a consciência fenomenal do estado interno ou de si próprio, *varredura interna* que reflete e lida com as informações internas, *metacognição* a qual permite estar consciente de um determinado estado e de estar consciente de estar consciente de um certo estado e assim por diante.

2.2.2 Problema mente-corpo

O *problema mente-corpo* nasceu do desejo por uma explicação do relacionamento entre o estado mental e o estado físico, da necessidade de explicar como a mente surge do cérebro. Esse problema engloba diversos questionamentos sobre a existência ou não de classes distintas para as entidades, processos ou propriedades mentais e físicas; buscando contrapor matéria e não matéria, físico e metafísico, objeto e sujeito.

Há tradicionalmente três teorias que buscam explicar o problema: a *mentalista*, a qual afirma que só há a mente que produz o mundo material e as suas sensações; a *materialista*, defensora de que a mente nada mais é que um processo físico; e a *dualista* a qual declara que a matéria e a mente são distintas e cada uma tem a sua própria existência (Franklin, 1995, cap. 2). Uma das principais vertentes no estudo do problema mente-corpo é o dualismo de Descartes que é apresentado a seguir.

Dualismo cartesiano

René Descartes (1596-1650) é usualmente identificado como o grande representante do dualismo (Young, 1990; Franklin, 1995; Seager, 2007; Zeman, 2008). Na sua obra *Discurso do Método*, publicada em 1637, Descartes argumenta que apesar de ser possível estar enganado sobre outras crenças, não é possível para ele se enganar em relação a ele ser uma "coisa que pensa":

> Decidi fazer de conta que todas as coisas que até então haviam entrado no meu espírito não eram mais corretas do que as ilusões de meus sonhos. Porém, logo em seguida, percebi que, ao mesmo tempo que eu queria pensar que tudo era falso, fazia-se necessário que eu, que pensava, fosse alguma coisa. E, ao notar que esta verdade: eu penso, logo existo, era tão sólida e tão correta que as mais extravagantes suposições dos céticos não seriam capazes de lhe causar abalo, julguei que podia considerá-la, sem escrúpulo algum, o primeiro princípio da filosofia que eu procurava (Descartes, 1637).

Diante dessa inferência considerada por ele como segura, Descartes conclui que corpo e mente são entidades completamente distintas:

> (...) compreendi, então, que eu era uma substância cuja essência ou natureza consiste apenas no pensar, e que, para ser, não necessita de lugar algum, nem depende de qualquer coisa material. De maneira que esse eu, ou seja, a alma, por causa da qual sou o que sou, é completamente distinta do corpo

> e, também, que é mais fácil de conhecer do que ele, e, mesmo que este nada fosse, ela não deixaria de ser tudo o que é (Descartes, 1637).

Diante de sua conclusão, Descartes tem sofrido uma série de críticas. Alguns sugerem que faltou coragem[4] a Descartes para que sua ciência terminasse com uma conclusão lógica e materialista. Entretanto, Descartes possuia argumentos para suportar a sua conclusão e que serviam de estratégias básicas antimaterialistas. Um exemplo é a atualmente conhecida como *Lei de Leibniz* que foi formulada por Gottfried Leibniz (1646-1716): *"se x tem uma propriedade que falta em y, então x e y não são idênticos"*. Assim, Descartes argumenta que a matéria possui um lugar no espaço, enquanto que a mente não pode ser medida em centímetros cúbicos (Seager, 2007).

O dualismo cartesiano clássico evoca Deus como o ponto de interação entre mente-corpo. Para Descartes, o ponto físico de interação, seria a glândula pinal. Interacionistas modernos entendem que a interação entre os eventos físicos e mentais ocorrem, mas ainda não explicam o fato em termos causais (Young, 1990).

As ideias de Descartes apesar de muito combatidas estão incorporadas nas mais diversas áreas do conhecimento. Na medicina, por exemplo, é comum encontrar as doenças agrupadas em *"somáticas"* e *"psicogênicas"*, uma distinção que assume que a psique é inorgânica (Zeman, 2008). Em outro exemplo, a ideia básica por trás da tecnologia de ficção científica da trilogia Matrix é também cartesiana: o que experenciamos não está diretamente relacionado com o estado do ambiente. Ao invés disso, o sistema nervoso é um mediador entre o ambiente. Nesse sentido, o estado de consciência atual e as experiências são provenientes do resultado da mediação cognitiva e sensorial provida pelo cérebro (Seager, 2007).

2.2.3 Qualia

Muito se sabe sobre o funcionamento fisiológico dos sentidos. A ciência tem diversas explicações para como se dá a visão ou a audição através dos mecanismos biológicos e fisiológicos do corpo humano. Mas, como acontecem os estados mentais provocados por uma entrada sensorial? Nesse contexto entram os *qualia*[5], que estão ligados à ideia de "como é experienciar algo", consagrada no notório artigo "Como é ser um morcego?" (Nagel, 1974) e também relacionados à "experiência qualitativa subjetiva" de (Chalmers, 1996).

Os qualia estão ligados às propriedades introspectivas das entradas sensoriais. Assim, os qualia da entrada visual de uma rosa incluiriam a experiência de vermelhidão e os qualia da entrada olfativa incluiriam a doçura de seu perfume (Amoroso, 2003).

Definir o que seriam os qualia não é uma tarefa tão simples. Em síntese são estados mentais do "como é ser", como "se sente", exemplificados pelas experiências subjetivas de sentir dor, ver o vermelho, cheirar uma rosa, beber um taça de vinho (Amoroso, 2003). Os estados mentais que possuem qualia são os ligados a: *experiência de percepção*, como ouvir o sino, sentir o licor, cheirar o ar do oceano, segurar uma pedra de gelo;

[4] É dito que o destino de Galileo influenciou Descartes, ou que, possivelmente, Descartes não queria prejudicar a igreja católica (Seager, 2007).

[5] plural de *quale*, um termo técnico introduzido inicialmente em (Lewis, 1929).

sensações corporais, como sentir fome, ter uma dor de estômago, sentir calor, sentir-se tonto; *sensações de reação, paixão ou emoção*, como sentir amor, medo, deleite, alegria, ciúme, arrependimento; e, por fim, a *sensação de humor*, como se sentir depressivo, calmo, chateado, tenso, miserável (Tye, 2007).

Apesar de ser normalmente associado aos aspectos introspectivos e fenomenais do funcionamento da mente, existem vários usos do termo qualia, principalmente na filosofia (Tye, 2007). Alguns pesquisadores como Daniel Dennett (ver seção 2.3.6), negam a existência deles. Entretanto, os qualia estão normalmente associados ao próprio entendimento da natureza da consciência e são intimamente relacionados ao problema mente-corpo (Chalmers, 1996; Tye, 2007).

2.2.4 O problema difícil da consciência

David Chalmers (Chalmers, 1995, 1996) divide o problema da consciência em *problema difícil da consciência* e *problema fácil da consciência*, divisão a partir da qual se pode fazer um paralelo com a diferenciação feita por (Searle, 1980) com relação à inteligência artificial[6].

O *problema fácil da consciência* está relacionado ao desempenho de funções cognitivas e pode ser resolvido pela descoberta de mecanismos neurais ou computacionais que o executem. Desse modo, ao se encontrar um mecanismo que resolvesse o problema de integração de informação pelo sistema cognitivo, um dos "problemas fáceis" estaria resolvido. Além desse, Chalmers (1995) relaciona ao *problema fácil da consciência* a habilidade de discriminar, categorizar e reagir aos estímulos do ambiente, a habilidade do sistema acessar e relatar seu próprio estado interno, o foco de atenção, o controle deliberado do comportamento, e a diferenciação entre o estado de sono e de vigília.

Já o *problema difícil da consciência* se refere ao problema da experiência, o aspecto subjetivo da consciência, aos *qualia*. Nesse último caso, para o problema ser resolvido seria necessário o entendimento de como se dá o aspecto da experiência consciente da consciência humana.

[6] Searle divide o estudo da inteligência artificial em *IA forte* e *IA fraca*:

> Eu acredito ser útil distinguir o que eu chamarei de *IA "forte"* da *IA "fraca"*. De acordo com a *IA fraca*, o principal valor do computador no estudo da mente é que ele serve como uma ferramenta muito poderosa. Por exemplo, ele nos possibilita formular e testar hipóteses de um modo preciso e rigoroso. Mas, de acordo com a *IA forte* o computador não é meramente uma ferramenta no estudo da mente, em vez disso, o computador programado de forma apropriada é uma mente no sentido que computadores e programas apropriados podem ser literalmente ditos que entendem e possuem estados cognitivos. Na *IA forte* como os computadores possuem estados cognitivos, os programas não são meramente ferramentas que permitem explicações psicológicas, ao invés disso, são eles mesmos as explicações (Searle, 1980).

2.3 Modelos de Consciência

2.3.1 Introdução

Diante de um cenário tão vasto de definições e conceitos sobre consciência, uma abordagem mais pragmática é a análise dos diversos modelos oferecidos pela literatura.

O objetivo dessa seção não é evidentemente trazer todos os modelos existentes, o que foge ao escopo desse trabalho. Ao invés disso, serão mostrados os principais estudos científicos sobre o tema nas diversas vertentes encontradas.

Os modelos podem ser divididos em: *orientados no espaço* - que são aqueles em que o lugar no cérebro realiza papel fundamental no modelo descrito, como o modelo de Edelman (seção 2.3.3) e o modelo de Damásio (seção 2.3.2); *orientados no tempo* - são os modelos que procuram características do cérebro (ou dos neurônios) que determinam quando acontece a consciência como inicialmente apresentado em (Crick & Koch, 1990) e continuado no trabalho de Llinás (seção 2.3.5); **modelos darwinianos** os quais se inspiram na teoria da evolução de Charles Darwin para criar os pontos chaves dos modelos no qual se destaca o trabalho de Dennett (seção 2.3.6); e por fim os *modelos quânticos*[7] os quais possuem um grau de abstração elevado trazendo explicações baseadas na mecânica quântica, como o eclético modelo de Penrose-Hameroff (seção 2.3.7).

2.3.2 António Damásio

O médico e neurocientista português António Damásio, baseado em evidências neurológicas empíricas apresentou a sua teoria em (Damasio, 1999). Através do estudo dos padrões mentais, Damásio procurou entender como o cérebro produz as "imagens de um objeto" ou como é obtido o "filme no cérebro". Além disso, procurou compreender também como é produzido o sentido de *self*[8], o que seria o senso de que existe alguém que é proprietário e observador desse filme. O sentido de *self* está ligado à capacidade do cérebro de possibilitar ao indivíduo o conhecimento, como organismo vivo e estabelecer os vínculos com os objetos e o mundo.

Em sua teoria, Damásio pressupõe uma forte ligação entre consciência e emoções. Mais que isso, ele defende que o estudo das emoções é fundamental na compreensão de um sistema nervoso integrado (Damasio, 1998). As emoções são um conjunto de reações químicas e neurais que auxiliam o organismo na preservação da vida. Apesar de terem sua expressão alterada pelo aprendizado e pela cultura, dependem de mecanismos cerebrais inatos que se desenvolveram de maneira evolutiva. Os mecanismos relacionados à emoção podem ser disparados de maneira inconsciente. Damásio categoriza as emoções em: *primárias ou universais* como alegria, medo, tristeza; *secundárias ou sociais* como ciúme, culpa, embaraço; e as *emoções de fundo* como bem-estar, calma ou tensão (Damasio, 2000).

Damásio também divide a consciência em *consciência central* e *consciência ampliada*. A primeira, mais simples e que não evolui com o passar do tempo, diz respeito

[7] Uma recente catalogação desses modelos pode ser encontrada em (Vannini, 2008).

[8] será mantida a palavra em inglês, como feita na tradução em (Damasio, 2000, p. 18)

ao momento presente e fornece o sentido de *self* do aqui e agora. Ela não está ligada à memória convencional, operacional, do raciocínio ou da linguagem, oferecendo apenas a percepção de que os pensamentos dos indivíduos são deles. Já a segunda, mais complexa, dá ao indivíduo a capacidade histórica, além de possuir um futuro esperado. Conta com vários níveis de organização e evolui durante a vida do organismo, dependente das memórias convencional e operacional. Essa consciência atinge seu ápice com o aparecimento da linguagem.

Dos dois tipos de consciência surgem dois tipos de *self*. O proveniente da consciência central é chamado *self central* que é recriado constantemente conforme a interação do cérebro com os objetos. O conceito mais tradicional de *self*, o qual possui um conjunto não transitório de fatos e modos únicos que caracterizam um indivíduo devido às suas *memórias autobiográficas* como seus gostos, aversões e experiências de vida, vem da consciência ampliada e é chamado por Damásio de *self autobiográfico*, o qual é construído sobre o *self central*.

Os dois níveis de consciência de Damásio têm como alicerce um conjunto de mecanismos cerebrais ligados à regulação da vida, que continuamente e inconscientemente buscam manter a estabilidade do sistema com a finalidade de sobrevivência. Esses mecanismos, precursores inconscientes dos níveis de *self*, são padrões neurais que mapeiam constantemente o estado do corpo do organismo. Damásio os chama de *proto-self*.

2.3.3 Neurodarwinismo de Edelman

Gerald Edelman, prêmio Nobel de fisiologia e medicina de 1972, propôs, em (Edelman, 1978), o Neurodarwinismo, também chamado de *Teoria de Seleção de Grupos Neurais* (TNGS)[9]. Inspirada nas teorias de evolução, a TNGS é uma teoria sobre o funcionamento do cérebro que tem sido relacionada ao fenômeno da consciência (Edelman, 1987; Tononi & Edelman, 1998; Seth & Baars, 2005; de Almeida & El-Hani, 2006).

Edelman (1992, p.82) considera a consciência como um processo multidimensional que emerge das interações entre corpo, cérebro e ambiente. Ele propõe que uma teoria de consciência deve trazer as explicações de como os processos psicológicos da consciência na mente estão ligados aos processos fisiológicos do cérebro.

Edelman defende que uma teoria de consciência não deve ser instrucionista, ou seja, relacionada com a noção de que o cérebro é um computador, com lógica, sinal de *clock* e um sistema de instruções simbólicas bem definidos[10]. Ao invés disso, propõe uma visão selecionista em que o cérebro é um sistema com uma enorme quantidade de circuitos diversos de neurônios que são gerados pelas sinapses, os chamados grupos neurais, os quais são as unidades básicas de seleção. A seleção acontece tanto durante o funcionamento do cérebro, como durante o desenvolvimento neural. A teoria de Edelman é baseada em três princípios básicos (ver figura 2.1) (Edelman, 1992).

[9] do inglês, *Theory of Neuronal Group Selection*.
[10] Alguns argumentos podem ser encontrados em (Edelman, 1992, p. 81-82).

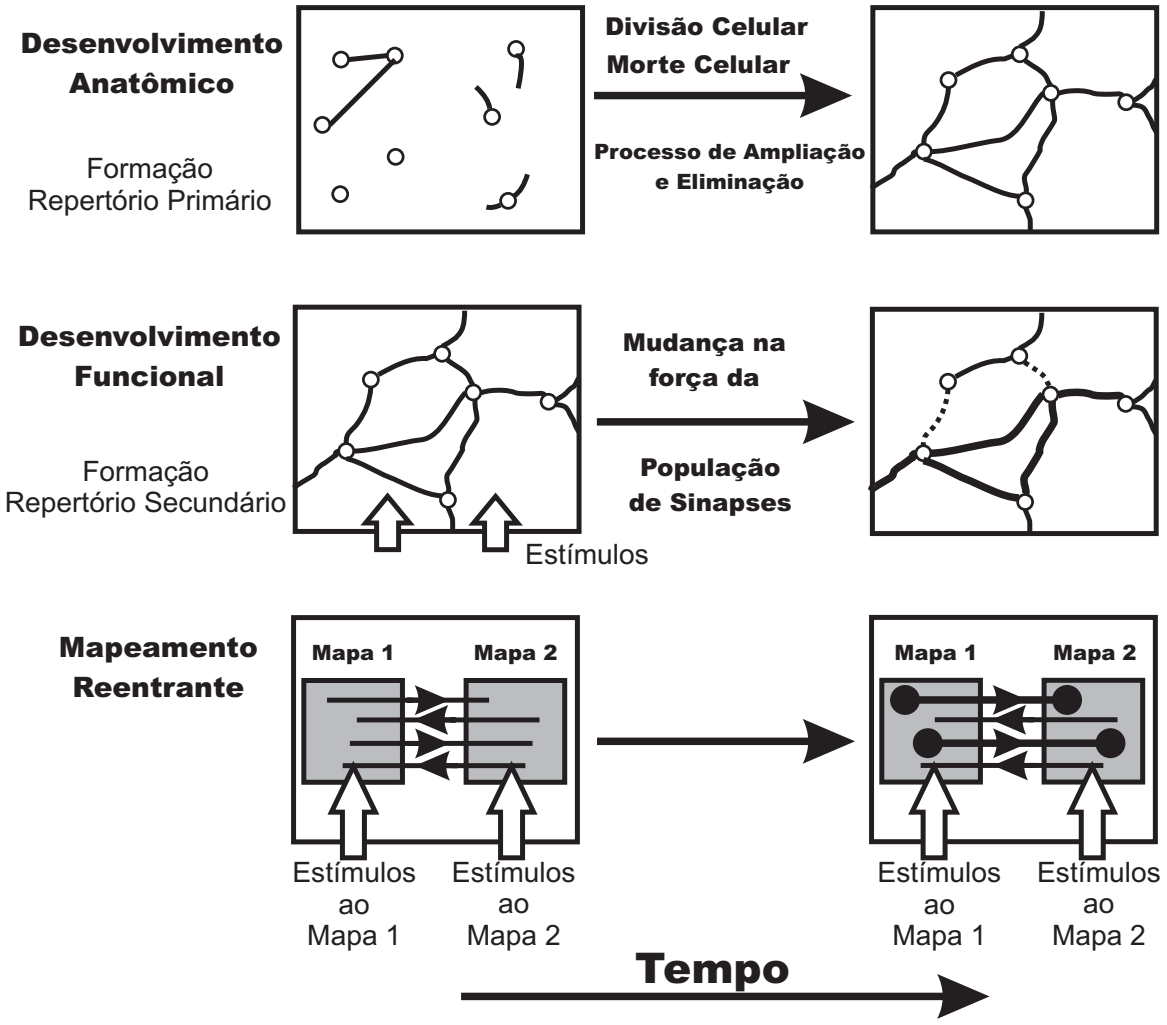


Figura 2.1: Princípios do Neurodarwinismo, adaptado de (Edelman, 1992, Fig. 9-1). A teoria de seleção de grupos neurais possui três princípios. Topo: Seleção no *desenvolvimento anatômico*, criam-se redes neurais variadas em cada indivíduo, chamado de *repertório primário*. Centro: seleção no *desenvolvimento funcional*, na qual ocorre o fortalecimento ou enfraquecimento da força de populações de sinapses através do comportamento e experiência, formando circuitos, o chamado *repertório secundário* de grupos neurais. As consequências do fortalecimento sináptico são indicadas por linhas mais escuras; o enfraquecimento, por linhas tracejadas. Base: Reentrância. As ligações entre os mapas ocorrem no tempo através da seleção paralela e da correlação dos mapas dos grupos neurais. Esses grupos recebem sinais separada e independentemente. Esse processo provê a base da categorização da percepção. Os pontos nas extremidades dos segmentos de conexões recíprocas indicam certo fortalecimento paralelo e relativamente simultâneo das sinapses nos caminhos reentrantes.

O primeiro princípio, diz que a seleção natural atua no desenvolvimento anatômico do cérebro em que ocorre a formação do *repertório primário* de grupos neurais. Nessa fase, variação e desenvolvimento se dão através da divisão, migração e morte seletiva

das células, e com o crescimento de axônios e dentritos. Esse primeiro repertório de neurônios é construído epigeneticamente[11], o que gera um alto nível de diversidade no sistema nervoso nascente.

O segundo princípio, afirma que, durante o desenvolvimento funcional do cérebro, a seleção atua na formação dos grupos neurais funcionais, o chamado *repertório secundário*. Esse repertório é formado através do fortalecimento ou enfraquecimento das sinapses no *repertório primário* que ocorrem pela experiência e comportamento. Assim, há uma amplificação seletiva das conexões funcionais do primeiro repertório criado na fase de desenvolvimento anatômico.

Após formados, os repertórios primário e secundário são capazes de desempenhar isoladamente um determinado número de funções. Entretanto, pelo terceiro princípio da teoria de Edelman, chamado de *reentrância*, diferentes áreas do cérebro, de forma coordenada, conseguem realizar novas funções (Edelman, 1992, p. 85). Para isso, os repertórios primários e secundários formam mapas. Esses mapas são interligados por conexões altamente recíprocas e paralelas. Edelman cita o exemplo do sistema visual dos macacos que possuem mais de trinta mapas diferentes, cada um com um determinado grau de segregação funcional (para orientação, cor, movimento, etc). Sinais reentrantes atravessam as conexões recíprocas, interligando os neurônios selecionados em um mapa aos de outro mapa. Isso possibilita que vários mapas sejam selecionados ao mesmo tempo. A correlação e a coordenação desses eventos de seleção são alcançados através dos sinais reentrantes e do fortalecimento das conexões entre os mapas em um determinado tempo. Desse modo, a *reentrância* também serve como um coordenador espaço-temporal e é um conceito chave para a ligação entre Neurodarwinismo e consciência (Edelman, 2003).

Tanto na formação do primeiro como do segundo repertório é verificado, como resultado da variação e seleção, um alto grau de redundância[12]. Isso acontece devido ao fato de vários grupos neurais, não necessariamente semelhantes na forma estrutural, produzirem um mesmo resultado particular. Essa redundância pode ser percebida, por exemplo, nas várias redes neurais distintas, que podem apresentar saídas motoras equivalentes; ou, ainda, pela notável capacidade de recuperação da maioria dos danos cerebrais, quando a execução das funções da parte danificada é realizada por outra população de neurônios (Edelman & Gally, 2001).

A seleção no cérebro é realizada através da *valoração*, que reflete a importância de um determinado evento para o organismo. O mecanismo de *valoração* pode ser comparado à pressão seletiva no desenvolvimento do organismo. Durante o desenvolvimento anatômico e funcional do cérebro, conforme as alterações sinápticas ocorrem, as manifestações somáticas são condicionadas ou limitadas pela *valoração* e aspectos relevantes do organismo são selecionados. Por exemplo, caso um indivíduo possua uma mão de determinada forma com uma propensão a agarrar de uma determinada maneira, irá favorecer as sinapses e padrões de atividade neural que conduzam a ações apropriadas.

[11] Não controlado genéticamente, mas por mecanismos fisicoquímicos de seleção que governam a morfologia da célula, seu movimento, sua diferenciação (célula cardíaca ou um neurônio). Algumas células morrem. Conexões se formam enquanto outras desaparecem. Ao final, nenhum animal possui conectividade neural idêntica, nem mesmo gêmeos com material genético perfeitamente igual.

[12] Também chamada na literatura de *degenerescência*.

Portanto, os valores são um referencial para o desenvolvimento e aperfeiçoamento da ação e categorização baseadas no cérebro (Edelman & Tononi, 2000).

Para criar um modelo de consciência, Edelman faz uma associação da *consciência primária* (sensorial) com as interações reentrantes entre categorização da percepção e memória. A *consciência primária* se refere à presença de uma cena multimodal da percepção e eventos motores no presente. Diferentes sinais que contribuem para a formação dessa cena (conectados ou não de forma causal) podem estar relacionados com o sistema de valores e aprendizados passados do animal. Nesse contexto, o animal pode alterar os seus comportamentos de modo adaptativo. Esses animais com *consciência primária* não são capazes de criar uma narrativa explícita (apesar de possuirem memória de longo prazo) e, no melhor dos casos, eles podem somente realizar um planejamento com a cena imediata no presente, possuindo seletividade e flexibilidade para lidar com ambientes complexos, o que é uma vantagem sobre os animais sem tal habilidade de planejamento (ver figura 2.2).

A segunda proposta de Edelman traz uma *consciência de alto nível*, a qual emergiu mais tarde na evolução dos seres vivos e pode ser observada em animais com habilidades semânticas, como os chimpanzés. Está presente na sua forma mais elaborada na espécie humana, que é única a apresentar o uso real de sintaxe e semântica de linguagem. Os indivíduos que possuem *consciência de alto nível* são capazes de quebrar os limites temporais da *consciência primária*. Através da linguagem, a dependência temporal de entradas do presente da consciência não é mais um limitante (Edelman, 2003).

Edelman & Tononi (2000), após novas evidências dos estudos da neurociência, apresentaram a *hipótese do núcleo dinâmico*, em que é sugerido que apenas uma pequena fração de grupos neurais contribui diretamente para o estado consciente. Assim, um grupo de neurônios pode contribuir diretamente para o estado consciente se fizer parte de um agrupamento funcional, que pode estar fisicamente distribuído, mas conectado por meio de circuitos de reentrada no sistema tálamo-cortical. Um agrupamento funcional é um grupo de elementos que tem forte interação entre eles, mas baixa dependência externa. Esse agrupamento funcional é também bastante diferenciado, exibindo um alto grau de complexidade.

2.3.4 Crick & Koch

Francis Crick, físico e bioquímico britânico, um dos idealizadores da famosa molécula de dupla hélice do ácido desoxirribonucléico (DNA), em 1953, e prêmio Nobel em 1962; e Christof Koch, neurocientista e professor de biologia e engenharia do Instituto Califórnia de Tecnologia, são pioneiros nos estudos dos correlatos neurais da consciência[13].

Observando a percepção e os processos visuais elementares como a percepção de cor ou de movimento, Crick e Koch buscaram explicar a experiência consciente em termos causais (Crick & Koch, 1990, 1998, 2003). O foco dado a mecanismos da consciência visual parte da premissa de que, a partir do momento em que existir uma explicação científica para um aspecto da consciência, ela poderá ser ampliada para outros mecanismos mais complexos (Crick & Koch, 1990).

[13] Um sistema neural N é um correlato neural da consciência (NCC) se o estado N está diretamente correlacionado com os estados da consciência (Chalmers, 2002).

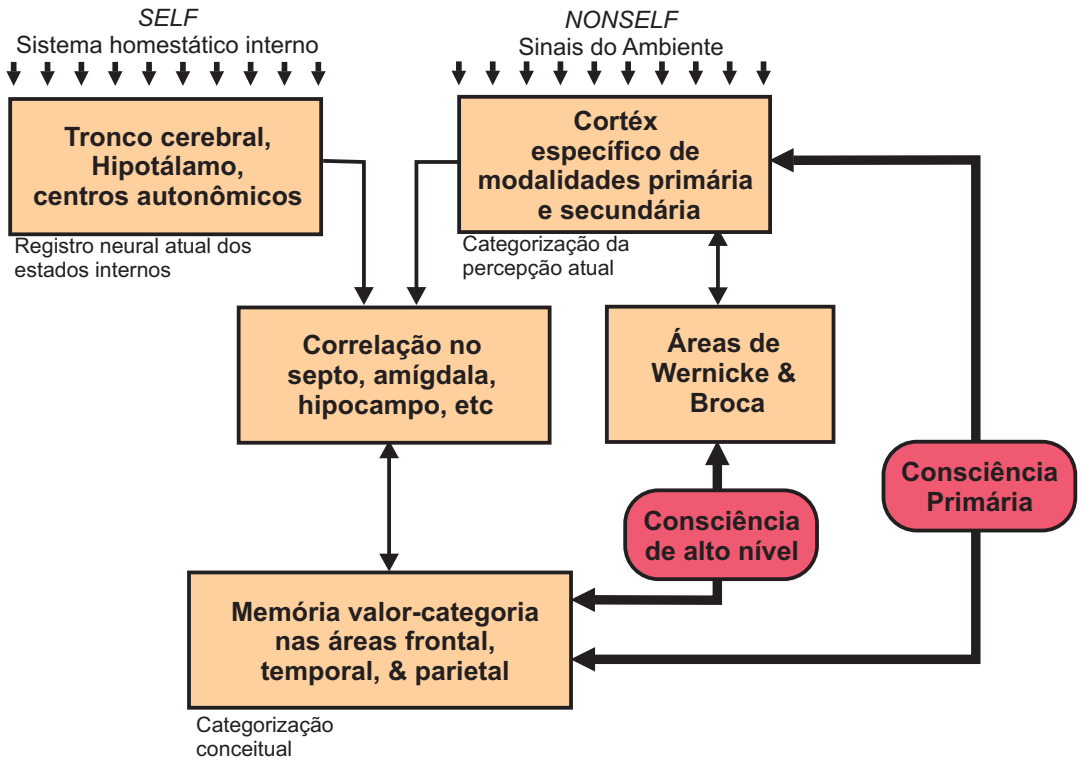


Figura 2.2: Modelo de consciência de Edelman, adaptado das fig 9.1 e 15.1 de Edelman & Tononi (2000). Os retângulos representam localidades, ou atividades em partes do cérebro. Sinais relacionados com os valores e sinais provenientes do mundo externo são correlacionados e produzem memórias valor-categoria. Essas memórias são ligadas por reentrância à categorização da percepção atual, o que resulta na consciência primária. A consciência de alto nível depende de *outras reentrâncias* entre a memória de valor-categoria e a categorização da percepção atual através das áreas que envolvem produção de linguagem e compreensão. Isso leva à explosão conceitual e permite que os conceitos de *"self"*, "presente", "passado" e "futuro" podem se conectar à consciência primária. Assim, é possível se "ter consciência da consciência."

Crick & Koch (2003) explicitamente abandonam a ideia de que os disparos síncronos de neurônios, as chamadas oscilações de 40Hz (Crick & Koch, 1990), são suficientes como correlatos neurais da consciência. Crick & Koch (2003) apresentam um modelo baseado no funcionamento do sistema cortical em que o córtex é uma rede neural interconectada. Muitos neurônios se relacionam de maneira excitatória ou inibitória, ocorrendo a formação de coalizões transientes de grupos de neurônios. Os neurônios de uma coalizão[14] reforçam um ao outro. Várias coalizões de diversos tamanhos podem aparecer ao mesmo tempo, podendo ser responsáveis pela experiência consciente,

[14]Uma coalizão é uma união entre entidades que se auxiliam em torno de um objetivo comum. A criação de várias coalizões implica na competição entre elas para alcançarem a consciência. As entidades assumem várias facetas dependendo da teoria: podem ser vistas como os grupos neuronais de (Edelman, 1992), neurônios em (Crick & Koch, 2003) e os processadores especializados de (Baars, 1997).

enquanto diversas outras são apenas formadas inconscientemente. Nas coalizões há o envolvimento de conexões recíprocas entre o córtex visual e outras áreas do cérebro. Para atingir a consciência, a atividade de um grupo neural deve ultrapassar um limiar de ativação (e permanecer por um certo tempo) o que acontece com a união de grupos em coalizões.

Figura 2.3: Hipótese dos instantâneos estáticos de Crick e Koch. Pela hipótese a percepção de movimento não é representada pela mudança da taxa de disparo dos neurônios relevantes, mas pelo disparo aproximadamente constante dos neurônios que representam o movimento. A figura faz uma analogia à possibilidade de uma foto estática sugerir movimento. Fonte: (Crick & Koch, 2003).

Crick & Koch (2003) defendem a hipótese de que a consciência visual está relacionada a uma série de *"instantâneos estáticos"*, "com movimento pintado sobre eles" (ver figura 2.3). Esses instantâneos não são uniformes variando em duração e na frequência. Por exemplo a duração de um instantâneo de forma pode não coincidir com um instantâneo de cor. A ideia dos instantâneos seria uma representação dinâmica de parte das coalizões bem sucedidas, uma vez que as coalizões não são estáticas.

2.3.5 Rodolfo Llinás

Rodolfo Llinás, professor da escola de medicina da Universidade de Nova Iorque estudou a consciência da percepção e o problema de integração da atividade consciente[15], o qual consiste no ato de mapear as unidades perceptivas responsáveis pelos

[15] em inglês, *binding problem*.

mecanismos que agrupam as diferentes componentes sensoriais em uma única imagem global, dado um conjunto de entradas geradas por algum objeto (e Silva *et al.* , 2003; Shadlen & Movshon, 1999).

Llinás (2003) defende que o cérebro é um sistema fechado capaz de gerar as suas próprias atividades, baseadas nas propriedades elétricas intrínsecas aos neurônios que o compõem e na conectividade entre eles. Além disso, para Llinás a base da consciência é a atividade interativa e recorrente na região tálamo-cortical.

Llinás observou que no córtex cerebral os neurônios, devido a suas propriedades elétricas, são capazes de oscilar a 40Hz (Llinás *et al.* , 1998). Além disso, através de magnetoencefalografia (MEG), foi possível identificar os pequenos campos magnéticos que são produzidos pelas correntes dos neurônios ativados em diferentes áreas do cérebro. Llinás percebeu que há uma atividade oscilatória coerente em todas as áreas do córtex cerebral, com uma frequência de 40 Hz durante a realização de atos cognitivos, sonhos ou estimulação sensorial, indicando uma causa comum no sistema de sincronização cortical. Assim, interações sinápticas recíprocas entre núcleos intralaminares do tálamo e o córtex seriam responsáveis por tal sincronização dadas as conexões tálamo-corticais (Llinás, 2003).

2.3.6 Daniel Dennett

Daniel Dennett, filósofo americano, escreveu vários livros sobre filosofia da mente (Dennett, 1969, 1991, 1995, 2005). Em sua teoria, a consciência é considerada um fenômeno físico e biológico como a reprodução ou o metabolismo, uma vez que o cérebro é um orgão como qualquer outro. Assim, ele nega a existência dos *qualia*, as características subjetivas da experiência consciente, rechaça o *dualismo cartesiano* e propõe um método de estudo da consciência através de ferramentas objetivas (Dennett, 1991).

Na sua teoria de consciência, Dennett utiliza o conceito de *meme*, criado por Richard Dawkins e apresentado inicialmente em (Dawkins, 1976) e mais recentemente também defendido em (Blackmore, 2003, 2005). Dawkins faz um paralelo da evolução cultural com a evolução genética: assim como gene é a unidade fundamental da hereditariedade e estaria sujeito às leis da seleção natural de Darwin, o *meme* seria a unidade fundamental da transmissão cultural. Dawkins (1976, p. 192), diz:

> 'Mimeme' vem de uma raíz grega apropriada, mas eu quero um monossílabo que soa como 'gene'. Espero que meus amigos classicistas me perdoem a abreviação de mimeme para *meme*. Se serve de consolo, poderia se pensar alternativamente sendo relacionado com 'memória', ou a palavra francesa *meme*. (...) Exemplos de memes são músicas, ideias, slogans, roupas, moda, os modos de fazer potes ou construir arcos. Como os genes se propagam no conjunto de genes pulando de corpo para corpo através de espermas ou ovos, os memes se propagam no conjunto de memes pulando de cérebro para cérebro através do processo que, de maneira geral, pode ser chamado imitação. Se um cientista ouve ou lê sobre uma boa ideia, ele a passa adiante para seus colegas e estudantes. Ele a menciona em seus artigos e em suas aulas. Se a ideia pega, diz-se que ela se propaga, espalhando-se de cérebro para cérebro.

Outro conceito importante para entender a teoria de Dennett é o de máquina virtual, bastante conhecido entre os cientistas da computação. Foi originalmente desenvolvido e utilizado em mainframes IBM na década de 60. Naquela época, cada máquina virtual simulava uma réplica física da máquina real dando a impressão aos usuários de que o sistema era de uso exclusivo (Goldberg, 1974). Nos arranjos tradicionais de máquinas virtuais, um software[16] roda sobre a máquina física, tendo controle total sobre o hardware. Esse software cria as máquinas virtuais que funcionam como se fossem máquinas independentes, podendo rodar o seu próprio sistema operacional (Sugerman *et al.* , 2001).

Apesar de Dennett considerar que o hardware seria a arquitetura paralela do cérebro, resultado de anos de evolução, e que a consciência humana é gerada por um imenso complexo de *memes* que criam uma máquina virtual sobre esse hardware, ele não segue o conceito tradicional dos cientistas da computação. Para Dennett cada aplicativo gera uma máquina virtual, por exemplo, um jogo ou programa editor de texto criariam uma máquina virtual especialista como um vídeo game ou um processador de texto. Assim, como uma máquina pode se transformar em um especialista, os seres humanos o fazem através da imitação: como a imitação de um braço quebrado, em que o indivíduo executa as mesmas limitações de movimento que teria caso estivesse com o braço engessado, apesar de fisicamente não ter nenhuma limitação (Dennett, 1991, p.209-218). Como essa imitação, outros memes possivelmente passam a fazer parte da vida de um indivíduo e esses geram máquinas virtuais que são responsáveis pelas várias atividades que o cérebro desempenha.

Dennett combate o *materialistmo cartesiano*, ao rejeitar a característica imaterial da "coisa pensante", que seria resquício do *dualismo cartesiano*[17]. Em (Dennett, 1991, p. 107) lê-se:

> Vamos chamar a ideia desse tal local central no cérebro de *materialismo cartesiano*, uma vez que essa é a concepção a que você chega quando se descarta o dualismo de Descarte mas se falha em remover a figura de um Teatro central (mas material). O Teatro é onde "tudo isso vem junto ". A glândula pineal poderia ser um candidato para esse *Teatro Cartesiano*, mas há outros que tem sido sugeridos: o cíngulo anterior, a formação reticular, vários lugares no lóbulo frontal. *Materialismo cartesiano* é a concepção de que há uma linha final crucial, ou uma fronteira, em algum lugar do cérebro, que marca um lugar onde a ordem de chegada é igual a ordem de "apresentação" na experiência porque *o que acontece lá* é o que você está consciente de.

Em contraste ao *teatro cartesiano*, Dennett sugere a *Teoria dos Múltiplos Rascunhos*:

> De acordo com o modelo de múltiplos rascunhos, todas as variedades de percepções - na verdade todas as variedades de pensamentos ou atividade

[16] chamado de *virtual machine monitor (VMM)*.
[17] Ver seção 2.2.2.

> mental - são executadas no cérebro em paralelo, por processos de múltiplos registros de interpretação e elaboração das entradas sensoriais. A informação entrante no sistema nervoso está em contínua "revisão editorial" (Dennett, 1991, p. 111).

Os processos de detecção ou discriminação de características em uma observação qualquer são realizados por porções especializadas do cérebro[18] de tal maneira que o conteúdo da informação não necessita ser enviado para um lugar mestre (como a pineal de Descartes). Essas informações são atualizadas constantemente pelos novos processamentos, podendo haver adições, incorporações, adendos, emendas, sobrescrições de conteúdo em qualquer ordem, ocorrendo "revisões editoriais" constantes.

Assim, a teoria de consciência de Dennett pode ser resumida da seguinte forma:

> A consciência humana é um enorme complexo de memes (ou mais precisamente, efeitos dos memes no cérebro) que podem ser melhor entendidos como a operação de uma máquina virtual "a la von Neumann" *implementada* em uma *arquitetura paralela* do cérebro que não foi projetada para nenhuma dessas atividades. Os poderes dessa *máquina virtual* melhoram os poderes subjacentes do *hardware* orgânico em que ela roda, mas ao mesmo tempo muitas das suas mais curiosas funcionalidades, e especialmente suas limitações, podem ser explicadas como um subproduto das *kludges*[19] que possibilitam esse curioso mas efetivo reuso do orgão existente para propósitos novos (Dennett, 1991, p. 210).

2.3.7 Penrose-Hameroff

Roger Penrose, físico matemático inglês e professor emérito da Universidade de Oxford, lançou o seu livro *The emperor's new mind*[20] associando o estudo da mente à mecânica quântica. O modelo de Penrose trouxe uma das descrições mais abstratas para o problema da consciência (Searle, 1997).

Em paralelo, Stuart Hameroff, médico anestesista e professor emérito da Universidade do Arizona, após seus estudos relacionados ao câncer, deu início a um grande interesse sobre a participação dos microtúbulos[21] na divisão celular. Hameroff (1987) especula que parte da explicação para o problema da consciência passa pelo entendimento do funcionamento nos microtúbulos nas células cerebrais. Apesar da teoria de Penrose tentar fazer a ligação do estudo da mente com a mecânica quântica, faltavam-lhe as explicações de como isso se dava no nível biológico, ou seja, faltava explicar o problema de integração da atividade consciente (*binding problem*).

[18] inspirado no trabalho Baars (1988). Ver seção 2.3.8.

[19] em inglês é uma gíria de computação para configurações de software e/ou hardware deselegantes e ineficientes que quando colocados juntos são bem sucedidos na resolução de um determinado problema ou ao desempenhar uma atividades particular. A palavra *"kludge"* vem da abreviação para *"klumsy, lame, ugly, dumb, but good enough"*, o que, em uma tradução livre, seria: *"descoordenado, imperfeito, feio, bobo, mas bom o suficiente."*

[20] ver (Penrose, 1989).

[21] um dos componentes do citoesqueleto celular.

Assim, da cooperação entre os dois cientistas nasceu a *teoria da redução objetiva orquestrada*[22], também conhecida como teoria *Orch OR*. Essa teoria surgiu da associação entre teoria geral da relatividade, lógica matemática, mecânica quântica, neurobiologia cognitiva e filosofia, o que daria início à Neurobiologia Cognitiva Computacional Relativística-Quântica (de Morais Ribeiro, 2001). O modelo *Orch OR* foi apresentado em (Penrose, 1994) e tem sido estudado em diversos outros trabalhos (Hameroff, 1998; Woolf & Hameroff, 2001; Hameroff *et al.* , 2002; Hagan *et al.* , 2002; Hameroff, 2007).

Penrose, em seu modelo de consciência, faz uso da *interpretação de Copenhagen*[23], uma das mais usuais interpretações da teoria quântica para o fenômeno do "colapso" ou da "redução" dos estados superpostos quânticos a um único estado clássico. Segundo a interpretação de Copenhagen, o colapso ocorre de forma aleatória após a medição do estado (chamada de *Redução Subjetiva* ***SR*** *ou* ***R***). Penrose (1989) adiciona a essa visão um novo ingrediente físico, a gravidade, lançando a chamada *Redução Objetiva* ***OR***. Pela interpretação de Penrose, sistemas quânticos coerentes podem sofrer o "autocolapso" quando alcançam um determinado limiar crítico de massa/tempo/energia, relacionado com a *gravidade quântica*. Assim, os estados colapsados não são necessariamente aleatórios mas refletem (de uma maneira não computacional clássica) uma computação quântica em um estado superposto coerente (Hameroff & Penrose, 2003).

Os microtúbulos são ótimos candidatos ao local em que ocorre a computação quântica da consciência. Eles se apresentam na forma de redes de polímeros protéicos (tubulina) no cito-esqueleto. Os microtúbulos dão forma aos neurônios, estabelecem e mantém as conexões sinápticas e têm papel chave nas tarefas essenciais desempenhadas por essas células como comunicação e processamento de informações. Alguns modelos teóricos apontam que, através dessa rede, os microtúbulos podem representar, propagar e processar a informação como em um "autômato celular" de sistemas computacionais (Hameroff & Penrose, 1996).

Penrose (1989), fazendo uso do teorema da incompletude matemática de Gödel, defende que a consciência humana não pode ser nem mesmo simulada em uma máquina de Turing clássica como um computador[24], sendo necessário um sistema capaz de produzir processamento quântico.

Em suma, segundo o modelo de Penrose-Hameroff, o fenômeno da consciência não

[22] tradução livre de *Orchestrated Objective Reduction*.

[23] Desenvolvida por Niels Bohr e Werner Heisenberg nos anos vinte, essa interpretação é fundamentada em três teses:

1. **Os resultados da mecânica quântica são não-determinísticos.** Por exemplo, no lançamento de um dado, mesmo acreditando que o processo é determinístico, usa-se probabilidades para prever o resultado porque não há informação suficiente do resultado final. Já na mecânica quântica, as previsões probabilísticas não são apenas o reflexo da falta de conhecimento.

2. **A Física é a ciência dos resultados de processos de medida.** Aquilo que não é medido não é contabilizado.

3. **A observação leva ao "colapso da função de onda".** Assim, em um sistema com um estado de muitas possibilidades (antes da medição), somente uma delas é escolhida após o processo de medição, o que é refletido pela função de onda instantaneamente modificada (Faye, 2008).

[24] Essa crítica ao estudo da consciência artificial será retomada na seção 3.5.

pode ser explicado utilizando-se a neurociência convencional. Ao invés disso, seria possível explicar por meio do fenômeno quântico de autocolapso instantâneo de estados superpostos, a *"redução objetiva orquestrada"* do pacote de onda provocada por efeitos gravitacionais instanciados nas atividades das redes neurais, que ocorrem nos microtúbulos do citoesqueleto celular.

O modelo de Penrose-Hameroff é alvo de muitas críticas. Primeiro: há várias pesquisas nas áreas de ciências cognitivas, psicologia e neurociências que são praticamente ignoradas por Penrose. Segundo: Penrose segue apenas argumentos lógicos e não apresenta nenhuma base empírica, vestígio ou medição do acontecimento dos fenômenos quânticos por ele defendidos. Por último, a teoria de Penrose não cita ou dá suporte aos eventos inconscientes que normalmente acompanham os modelos de consciência mais sofisticados (Baars, 1995; Searle, 1997, cap. 4).

2.3.8 Teoria do *Workspace* Global

Bernard Baars, pesquisador de ciências cognitivas do Instituto de Neurociências da Califórnia, desenvolveu a *teoria do workspace*[25] *global*[26] (Baars, 1988, 1997, 2002), inspirada na psicologia e fundamentada em testes empíricos das ciências cognitivas e da neurociência.

Além de representar um passo importante na direção de uma descrição concreta de como a consciência humana funciona, essa teoria tem sido bastante utilizada na criação de modelos computacionais da consciência (Franklin & Graesser, 1999; Bogner, 1999; Dehaene *et al.* , 1998, 2003; Dehaene & Changeux, 2005; Moura, 2006; Negatu, 2006; Shanahan, 2006; Dubois, 2007a; Shanahan & Connor, 2008).

A teoria de Baars foi desenvolvida através de um conjunto de estudos que buscavam comparar e estudar as características dos casos em que havia consciência, contrastando-as com as propriedades dos casos em que a consciência não estava presente. Essas *análises por contraste*[27] oferecem importante fundamentação empírica e são utilizadas por vários pesquisadores[28] para estudar fenômenos conscientes e inconscientes. Apesar disso, alguns preferem evitar a palavra "consciência" e falar em "controle estratégico versus controle automático" ou "memória imediata versus memória de longo prazo", dentre outros termos que podem ser observados na tabela 2.1 (Baars, 2003).

Componentes da Teoria do Workspace Global

Na teoria do workspace global, a consciência tem função integradora e mobilizante (ver figura 2.5). Ela cria um acesso global que auxilia no recrutamento e integração de

[25] A palavra *"workspace"* pode ser traduzida por *"área de trabalho"* ou *"espaço de trabalho"*. Entretanto, preferiu-se manter a palavra original em inglês devido a nenhuma das traduções conter a carga semântica adequada da palavra em inglês.

[26] em inglês *Global Workspace Theory*.

[27] Por exemplo, entre o conteúdo consciente e inconsciente da memória, entre as formas de lesão cerebral que seletivamente impactam os processos conscientes e as que não impactam, entre o estado de vigília, de sono profundo e de coma.

[28] principalmente neurocientistas e estudiosos das ciências cognitivas.

Tabela 2.1: Alguns tipos de estudo de fenômenos consciente e insconsciente (reprodução parcial de Baars (2003))

Consciente	Inconsciente
Cognição explícita	Cognição implícita
Aprendizado intencional	Aprendizado incidental
Inferências explícitas	Inferências automáticas
Memória imediata	Memória de longo prazo
Memória episódica (autobiográfica)	Memória semântica (conhecimento conceitual)
Memória declarativa (fatos, etc)	Memória procedural (habilidades, etc)
Eventos novos, informativos e significantes	Eventos insignificantes, rotineiros, previsíveis
Tarefas elaboradas	Tarefas espontâneas/automáticas
Controle estratégico	Controle automático

diversas funções separadas e independentes do cérebro e também coleções inconscientes de conhecimentos.

Baars não cria uma teoria completamente inovadora, mas agrupa diversas hipóteses isoladas. Atenção, seleção de ação, automatização, aprendizado, emoções, metacognição, entre outras funções cognitivas são explicadas por Baars através de componentes básicos de sua teoria como: *processador especializado*, *coalizão*, *contexto* e *workspace global*. Esses componentes podem ser vistos na figura 2.4 e serão descritos a seguir.

Processadores e Coalizões

Baars (1988), fundamentado em evidências psicológicas e neurofisiológicas, sugere que o sistema nervoso contém vários *processadores especializados*, que são autônomos e operam, na sua maioria, de forma inconsciente. Cada processador pode ser definido como uma coleção unitária de processos. Eles podem trabalhar em conjunto a serviço de uma determinada função, como detecção de uma característica ou a execução de uma ação primitiva. Além disso, esses *processadores especializados* são bastante eficientes, trabalham em paralelo, em alta velocidade, e cometem poucos erros. O paralelismo permite a criação de um sistema de altíssima capacidade como o sistema nervoso central.

Os *processadores especializados* podem cooperar entre si formando *coalizões*, a fim de desempenhar um grande número de tarefas. As coalizões podem desempenhar atividades inconscientes e automatizadas como acontece nas atividades de respiração e de circulação sanguínea. Também podem ocupar a consciência quando tratam de situações novas ou quando as atividades automatizadas passam a não gerar o resultado esperado. As coalizões têm um caráter recursivo, podendo ser compostas por outras coalizões.

Um processador tem o potencial de trazer o seu conteúdo para a consciência através do seu *nível de ativação*. Quando um processador traz informações novas, tem nível de ativação maior do que aqueles que trazem informações de rotina. Os níveis de ativação dos processadores contribuem para o nível de ativação da coalizão que participam. Assim, coalizões que passam a ter um papel ativo na resolução de ambiguidades ou

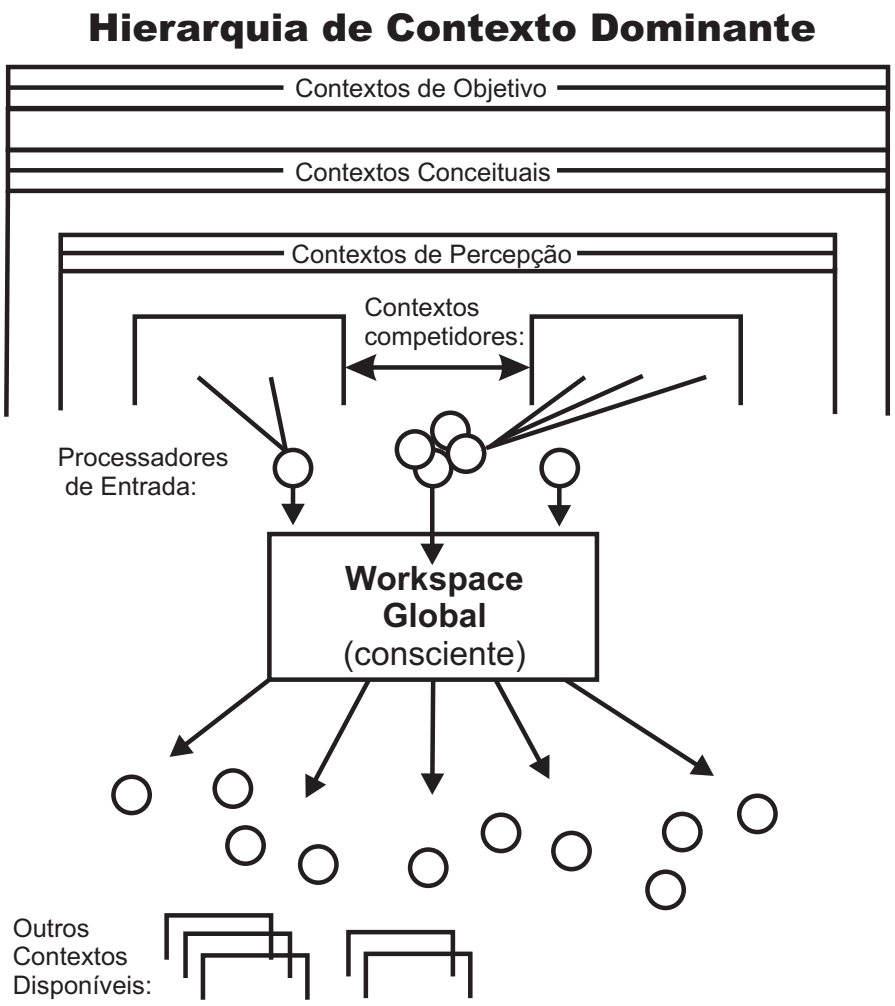


Figura 2.4: Componentes da teoria do *workspace* global. Adaptado de (Baars, 1988, fig. 4.5)

conflitos tendem a possuir um nível de ativação elevado, o que aumenta a chance de ter acesso à consciência[29]. Por outro lado, coalizões que estejam executando atividades rotineiras tendem a ter um nível de ativação baixo, mantendo-se inconscientes.

Contextos

Contextos são coalizões de processadores inconscientes, relativamente estáveis no tempo. Os contextos, mesmo sendo em sua maioria inconscientes, afetam a experiência consciente diretamente, evocando ou modelando mensagens globais ou conteúdos conscientes. Os contextos representam conhecimento prévio ou pressuposições que ao presenciarem uma realidade discrepante tendem a levá-la à consciência. Novos contextos podem ser aprendidos, permitindo que a realidade seja melhor percebida.

Outros termos da ciência cognitiva estão relacionados com o conceito de contexto, como "estruturas ativas de conhecimento", "representações mentais", "redes semânticas",

[29] Um nível de ativação alto é condição necessária mas não suficiente para ter acesso à consciência.

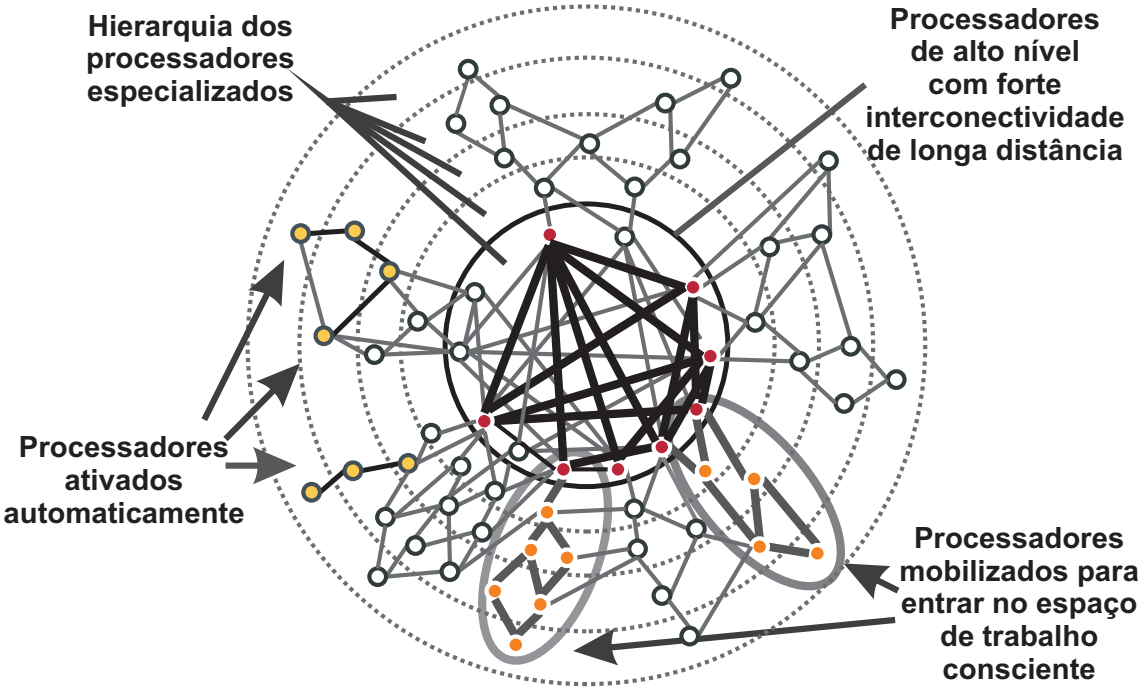


Figura 2.5: Ativação de processadores na teoria do *workspace* global. Processadores ativados são simbolizados por pontos coloridos (amarelos, laranjas, vermelhos) e linhas escuras. No centro há a coalizão central na consciência (formada pelos processadores em vermelho). O conteúdo dessa coalizão gera uma atividade de excitação de alguns outros processadores (laranjas) os quais criam outras coalizões, que entram na disputa para entrar na consciência. Alguns processadores (amarelos) podem ser ativados automaticamente (ou "inconscientemente").

"quadros", "esquemas", "scripts", "planos", "expectativas", dentre outros termos ligados à representação do conhecimento (Baars, 1988). Apesar da similaridade, os contextos de Baars não são meramente representações mentais, mas representações inconscientes que influenciam outras representações conscientes. Há quatro tipos de contexto: *contexto de objetivo*, *contexto de percepção*, *contexto conceitual* e *contexto cultural* (ver figura 2.4).

Os *contextos de objetivo* representam estados futuros desejados do sistema. Eles limitam as concepções conscientes ou intenções de resultado futuro, uma vez que reduzem o escopo ao uso de processos - ou de conteúdo consciente - que levam ao estado futuro pretendido. Além disso, os contextos de objetivo podem ser hierarquizados com objetivos e sub-objetivos. Por exemplo, nesse momento a experiência consciente do leitor está possivelmente ligada à meta de ler essa frase e para isso a consciência está restrita a processos que auxiliam nesse resultado. Ao terminá-la, ou ao finalizar a seção, fica evidente que contextos de objetivo local existem e que há uma hierarquia complexa de objetivos em que ler aquela frase era apenas um simples sub-objetivo local. Nesse sentido, para que um objetivo de mais alto nível seja alcançado, é necessário que os sub-objetivos também sejam atingidos.

O *contexto de percepção* armazena pressuposições as quais restringem as experiências

relacionadas à percepção. Por exemplo, ao olhar uma imagem de cabeça para baixo da superfície da lua as crateras se transformam em montes uma vez que há a premissa de que a luz vem de cima o que faz com que as concavidades se tornem convexidades na foto invertida (como na figura 2.6).

Figura 2.6: Exemplo de contexto de percepção. A esquerda, a imagem de crateras na superfície da lua. A direita, a mesma imagem apenas rotacionada em 180^o: as crateras se tornam montes devido ao contexto de percepção da fonte de luz.

O *contexto conceitual* limita a consciência em relação aos conceitos abstratos e está ligado à noção de pressuposições estáveis que se tornam, em geral, inconscientes. Nesse caso, certas conjecturas podem se tornar fortes crenças, de tal maneira que algumas pessoas passam a ignorar a possibilidade de aparecer alternativas a essas pressuposições estáveis. Mesmo nos estudos científicos na chegada de uma nova teoria isso pode acontecer, como por exemplo, a forte resistência sofrida tanto por Einsten, com a teoria da relatividade, como por Darwin, com a teoria da evolução (Baars, 1988).

O *contexto cultural* se refere às restrições ao acesso consciente a interações sociais e pode ser observado, por exemplo, em uma típica aula de engenharia. Ambos, estudantes universitários e docentes, sabem o comportamento esperado em uma sala de aula, embora nem sempre isso seja consciente. Se em alguma aula o professor começasse a cantar, os estudantes se tornariam conscientes de que o professor está exibindo um comportamento anormal. Essa atitude do professor se oporia ao contexto cultural prevalecente. Em outro exemplo, problemas ligados ao contexto cultural podem ser facilmente encontrados quando pessoas estão tendo uma experiência fora do seu país, em que modelos novos de relações sociais são constantemente apresentados e conceitos inconscientes vem à tona devido ao choque cultural.

Workspace Global

O *workspace global* é o ponto central da teoria de Baars pois é a estrutura que suporta a experiência consciente. É uma memória de trabalho que media a troca de informações e as novas interações entre processadores.

Esse componente tem processamento serial com capacidade muito menor, se comparada com a do sistema criado pelos numerosos processadores inconscientes agindo em paralelo.

Devido a sua característica serial, uma única coalizão pode estar no *workspace* global. Isso provoca uma competição entre elas através dos seus níveis de ativação. A coalizão que chega ao *workspace* global ganha o direito de transmitir o seu conteúdo (consciente) a todos os outros processadores.

A metáfora do teatro interativo

Para facilitar o entendimento da dinâmica de sua teoria, Baars (1997) utiliza a *"metáfora do teatro interativo"*[30] para explicar o funcionamento da teoria do espaço global. É importante ressaltar que a metáfora do sala de teatro é somente uma maneira conveniente de lembrar as funcionalidades básicas da teoria do *workspace* global, em que é desenvolvido um conjunto de hipóteses que podem ser testadas (Baars, 2002).

Na sala de teatro interativo de Baars, é possível encontrar o palco, um holofote que orienta a atenção do público, atores (ou *processadores*) que competem pelo holofote, um público de processadores especializados e, nos bastidores, um conjunto de auxiliares (como os cenógrafos, figurinistas, sonoplastas, etc), que influenciam diretamente o que está sob o holofote.

No palco, ocorrem várias atividades que possuem diversas informações, mas somente o que ocorre sob a luz do holofote é consciente. Com o desenrolar da peça, mais e mais eventos ocorrem no palco e os auxiliares ou *processadores inconscientes* influenciam os eventos conscientes. O holofote seleciona os atores mais importantes no palco. Sob o holofote os atores desempenham suas ações e têm a sua mensagem transmitida para toda a plateia e para os operadores que estão nos bastidores. Várias atividades acontecem inconscientemente na plateia, nos bastidores e na parte não iluminada do palco.

O teatro de Baars é um teatro interativo. Quando os espectadores se identificam com a mensagem sendo transmitida sob o holofote, eles podem subir no palco e realizar sua própria performance em conjunto com os atores do holofote, ou mesmo na parte escura do palco. Nesse último caso, eles podem se tornar o foco principal do público caso o holofote seja apontado para eles.

Na teoria do *workspace* global, o conteúdo consciente é o resultado do processamento da coalizão mais importante, a qual está sob o brilho do holofote, caracterizando a limitação de capacidade e a característica serial. Essa coalizão que está no centro das atenções desempenha tarefas relacionadas a ítens ambíguos, conflitantes ou situações novas. Os processadores inconscientes (da audiência ou dos bastidores), quando encontram uma mensagem global relevante[31], executam a sua função específica.

2.4 Conclusão

Esse capítulo efetuou um sumário dos principais estudos de consciência, bem como os conceitos importantes nos estudos do tema. Como se pode notar algumas ideias

[30] É importante ressaltar que a *metáfora do teatrointerativo* de Baars difere do *teatro cartesiano* (ver seção 2.3.6), uma vez que não segue a linha de que há um único ponto central, como a glândula pineal. Ao invés disso, nessa metáfora Baars sugere uma sala de teatro real, a qual serve para auxiliar o entendimento da teoria do *workspace* global (Baars, 1997, p. viii-ix).

[31] A decisão se uma mensagem é relevante ou não é realizada localmente por cada processador.

aparecem nas diversas teorias, inclusive podendo ser encontrados artigos na literatura que buscam explorar essa convergência como em (Baars *et al.* , 1998).

Diante dos modelos apresentados nesse trabalho podemos citar as semelhanças entre os autores Damásio e Edelman na divisão da consciência: ambos dividem em duas, uma ligada ao presente e outra mais robusta que quebra as barreiras do momento atual e recapitula informações do passado contribuindo para uma habilidade ímpar de planejamento. Ambos também fazem a associação dessa consciência mais elaborada com o aparecimento da linguagem.

Edelman com os grupos neurais[32], Baars e Crick & Kock com as coalizões, Dennett com os processos e Baars com os processadores especializados, defendem que certos grupos realizam atividades específicas na construção da consciência.

Além disso, Llinás e Edelman ressaltam a importância das interações tálamo-corticais para o surgimento da consciência e, Baars e Penrose-Hameroff argumentam sobre a existência de um *nível de ativação* mesmo que em escopos bem diferentes, um nos processadores específicos (ou grupo de neurônios) e o outro no nível quântico.

Por último, é possível verificar uma conexão entre o conceito de reentrância de Edelman, que liga regiões distintas do cérebro propagando sinais de uma para a outra, com a ideia de broadcast de Baars, pela qual são transmitidos o resultado da coalizão vencedora. A ideia de ampla divulgação de informação também pode ser observada na A-Consciência de Block, em que uma representação só é considerada A-consciente se adequadamente divulgada e disponibilizada para uso.

[32]Se desconsideramos as características de redundância.

Capítulo 3

Consciência Artificial

> Things are only impossible until they're not.
>
> Jean-Luc Picard

3.1 Introdução

Esse capítulo traz os estudos recentes de consciência artificial (ou modelos computacionais de consciência), complementando a parte teórica de modelos gerais de consciência desenvolvida no Capítulo 2. Assim como nos estudos das outras áreas de conhecimento, na engenharia há também um crescente número de pesquisas e abordagens das mais diversas.

Na seção 3.2, são discutidas as principais motivações do estudo de consciência artificial. A seção 3.3, apresenta as principais arquiteturas cognitivas ligadas ao estudo de consciência artificial. Essas arquiteturas são concebidas para resolverem problemas genéricos e por isso foram foco da pesquisa bibliográfica realizada[1]. A seção 3.4, traz alguns estudos de consciência artificial relacionados a problemas específicos[2]. Por fim, a seção 3.5 aborda as objeções mais comuns encontradas na literatura em relação ao estudo de consciência artificial.

3.2 Por que estudar consciência artificial?

O crescimento do interesse nos estudos de consciência em geral e, principalmente, nas pesquisas envolvendo consciência artificial se dá devido a interesses da robótica e das ciências cognitivas, como será verificado através dos exemplos das seções 3.3 e 3.4. De uma maneira geral, Sanz *et al.* (2007) agrupa as três principais motivações relacionadas à área de consciência artificial: criar artefatos mais parecidos com seres

[1] Esse estudo de arquiteturas cognitivas será completado no capítulo 4, descrevendo mais detalhadamente a arquitetura Baars-Franklin, a qual será a base dos experimentos realizados nesse trabalho.

[2] Para uma extensa lista de trabalhos relacionados à consciência artificial ver (Gamez, 2008a, Sec. 3.5).

humanos, estudar sistemas naturais através de modelos computacionais e criar máquinas mais eficazes; as quais são discutidas a seguir.

Criação de artefatos mais parecidos com os seres humanos

Vários pesquisadores (ver seção 2.3) apontam a consciência como um diferencial dos seres humanos para lidar com situações complexas e diversificadas.

Através dos estudos de cognição artificial procura-se compreender melhor as vantagens trazidas pelos mecanismos de consciência, emoção e afeto, experiência, imaginação; e adicionar esses benefícios a robôs ou agentes de software. Uma das principais estratégias da robótica é criar comportamentos semelhantes aos comportamentos humanos por meio da adição desses mecanismos a robôs e agentes de software. Isso potencialmente auxiliaria na interação homem-máquina (Sanz *et al.* , 2007).

Dubois (2007a) realizou um estudo interessante em que se produziu um tutor "consciente" no contexto de treinamento de astronautas na utilização do sistema robótico de manutenção da Estação Espacial Internacional. A interação na tutoria da máquina com os aprendizes exige uma série de sutilezas para que o agente seja percebido como um bom tutor e um agradável companheiro. Assim, um agente tutor necessita de habilidades integradas e não somente de ter embutido conhecimento específico do assunto. O que parece produzir resultado é realizar o acompanhamento do conhecimento e concepções errôneas dos estudantes e, de forma adaptativa, responder às lacunas de aprendizado apropriadamente.

O agente desenvolvido permitiu grande flexibilidade de se adaptar ao ambiente graças ao seu mecanismo de atenção e consciência funcional. Isso foi o que permitiu que ações criativas e não automáticas fossem requisitadas quando os processos automáticos não traziam os resultados esperados (Dubois, 2007b).

Estudo de sistemas naturais através de modelos computacionais

No estudo de sistemas naturais, a verificação de teorias e hipóteses pode não ser uma tarefa simples. Especificamente no estudo de ciências cognitivas, o modo analítico de investigação nem sempre é suficiente para confirmar os diversos modelos propostos.

Para suprir essa necessidade, uma das principais estratégias de pesquisa em ciências cognitivas é a simulação computacional. Na área de consciência, alguns cientistas, como Nagel, relatam a inacessibilidade dos estados conscientes por um terceiro. Nesse sentido, o campo de consciência artificial se torna ainda mais relevante na compreensão dos mecanismos que fazem emergir a consciência humana.

Vários pesquisadores têm realizado trabalhos empíricos seguindo a linha de modelagem e simulação de mecanismos da consciência. Dehaene e Shanahan são exemplos de pesquisadores dessa linha. Em (Dehaene *et al.* , 2003; Dehaene & Changeux, 2005; Shanahan, 2007; Shanahan & Connor, 2008) são estudadas várias tarefas de cognição e percepção, através de simulações neurais para averiguar a influência do *workspace* global[3] nessas atividades.

[3] ver seção 2.3.8.

Criação de máquinas mais eficazes

Por último, diante de um contexto de busca incessante por tecnologias de controle inteligentes, mecanismos de consciência podem ser utilizados para que os agentes possam lidar com situações cada vez mais complexas. Assim, seria possível criar máquinas mais eficazes.

Nessa linha de pesquisa, pode-se citar o Cicerobot (ver seção 3.4), um robô guia de um museu na Itália capaz de realizar trajetos gerados a partir da interação com os visitantes.

3.3 Arquiteturas Cognitivas

3.3.1 Sistema Cognitivo Neural de Haikonen

Pentti Haikonen, pesquisador de tecnologias cognitivas do Centro de Pesquisas da Nokia na Finlândia, desenvolveu um sistema cognitivo baseado em redes neurais (Haikonen, 2000a,b, 2003). Haikonen propõe que uma mente artificial deve ser capaz de replicar os processos da mente humana como imagens mentais[4], conversação interna[5] e sensações, além de conter funções cognitivas como: introspecção, percepção, atenção, casamento de padrões, detecção de novidades, aprendizado, memória, raciocínio, planejamento, emoções e motivação (Haikonen, 2000b) .

No sistema proposto, ao invés de utilizar sistemas especializados em blocos explicitamente programados, Haikonen desenvolveu uma rede neural recorrente em blocos que podem ser generalizados para produzir uma máquina cognitiva. A arquitetura de Haikonen foi simulada em computador e possui entradas textuais e visuais. Nos testes realizados o sistema pode aprender a reconhecer as figuras, o significado concreto de uma palavra, palavras abstratas e sintaxes rudimentares. Além disso, o sistema foi capaz de aprender novas figuras por descrição verbal, detectar afirmações e contradições e deduzir as propriedades de um determinado objeto, evocando as suas imagens mentais (Haikonen, 2000a).

A sua arquitetura cognitiva (ver fig. 3.2) possui módulos de percepção, como os de processo visual, em um grande número de sinais (on/off) que transmitem a informação sobre as características dos estímulos. As entidades percebidas são representadas usando uma combinação desses sinais (um importante aspecto da teoria de consciência de Haikonen). Para Haikonen, a cognição acontece quando há associação de significados às percepções através do uso de símbolos, da sua manipulação, raciocínio, geração de resposta e de linguagem. Os subsistemas da arquitetura são detalhados a seguir.

Sistema Linguístico

O sistema linguístico percebe palavras de entrada, associa percepções de outros módulos às palavras, ou percepções de palavras a outras palavras, através da sua representação interna, e habilita a execução da narrativa interna. As palavras são representadas

[4] em inglês *inner imagery*.
[5] em inglês *inner speech*.

Figura 3.1: Neurônio Associativo de Haikonen. Adaptado de (Haikonen, 2000a, Fig. 3.1). O neurônio associativo tem uma entrada principal s, um certo número de entradas associativas a_i, um sinal de controle de limiar de aprendizado sináptico THs, um sinal de controle limiar bidirecional da saída do neurônio TH e sinais de saída: so, m, mm, n. m, mm e n, representam, respectivamente, sinais de: casamento de padrão (em inglês, *match*), falta de combinação (em inglês, *mismatch*) e detecção de novidades. O número de entradas associativas a_i não é limitado ou fixo, uma vez que seus pesos sinápticos não são ajustados entre si. Todos os valores de entrada ou saída podem assumir somente valores entre zero e um.

por sinais de letras distribuídos de modo que cada letra é representada por um sinal (*on/off*). Por princípio, como as palavras são sequências temporais de fonemas ou letras, as representações dos sinais de letras são transformadas em uma forma paralela de tal modo que as letras de uma palavra representada estão em posições fixas e estão disponíveis simultaneamente (Haikonen, 2003, Cap. 16).

Sistema Visual

O sistema visual percebe objetos visuais através dos padrões de cor, forma e tamanho. Ele associa objetos visuais a outros objetos visuais, evoca representações de objetos visuais internamente, através da percepção de outros módulos, e percebe esses objetos através do *feedback* associativo, o que habilita o fluxo de imaginação interna. Esse sistema também suporta o processo de significação de palavras concretas: ao perceber as palavras com significado visual, recupera a representação interna do respectivo objeto visual.

O sistema visual pré-processa imagens em sinais distribuídos de características para forma, cor e tamanho. Cada uma dessas características tem seu próprio loop reentrante de percepção e resposta. Imagens não são reconstruídas em ponto algum, uma vez que o sistema não opera internamente com imagens reais. A associação entre características e entidades reconhecidas são realizadas automaticamente através dos sinais associativos de ganho e de limiar (Haikonen, 2003, Cap. 13).

Sistema de Foco de Atenção Visual

Como somente um pequeno número de entidades pode ser ativamente processado em um determinado tempo, há um sistema para selecionar a atenção. O sistema de foco de atenção visual controla a posição do foco e temporariamente associa objetos visuais a suas posições. A posição do foco de atenção visual é determinada pela mudança visual, detectada por um pré-processador visual, ou através de uma chamada interna explícita para que essa mudança ocorra.

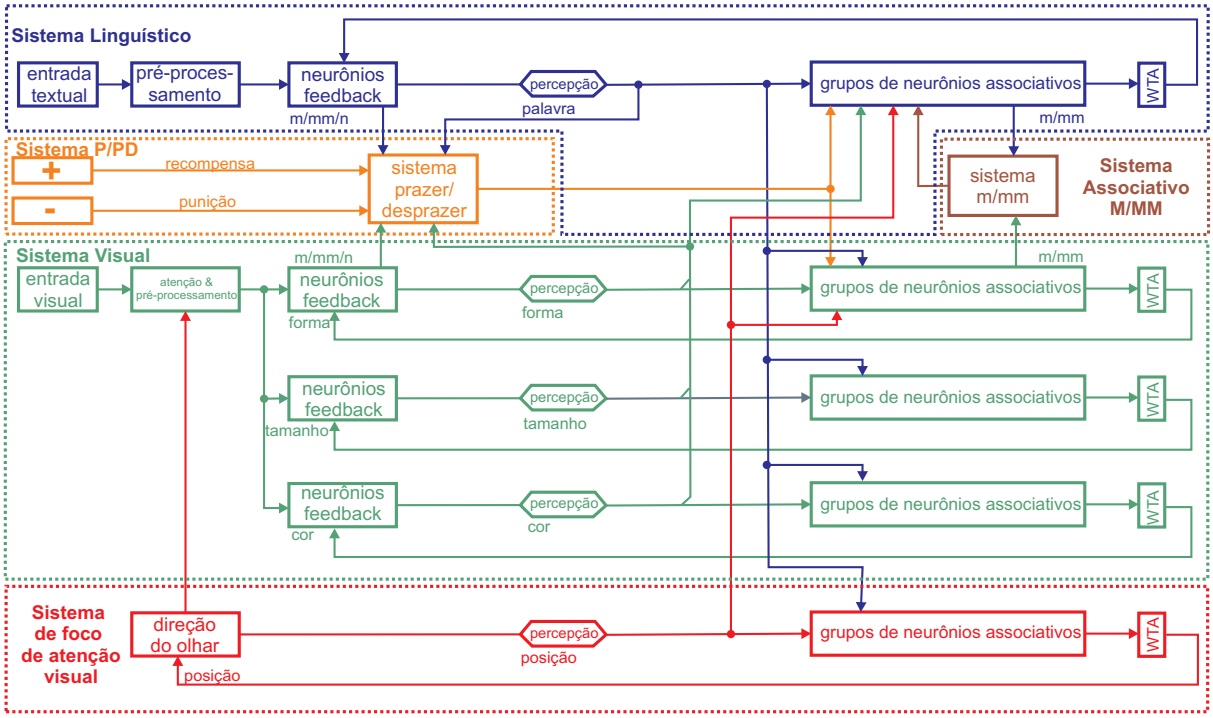


Figura 3.2: Arquitetura do Sistema Cognitivo de Haikonen. Adaptado de (Haikonen, 2000a, Fig. 5.1).

Sistema de Prazer/Desprazer

O sistema de prazer e desprazer (*Sistema P/DP*) associa um sentido de bom/ruim às percepções e guia o julgamento, motivação e atenção. Há dois tipos de entradas para esse sistema: os chamados *estados-casados*, vindos dos neurônios de *feedback* que geram prazer e são sinais de recompensa, e os *estados não casados*, que geram desprazer e são sinais externos de punição.

Prazer ou desprazer relacionados com recompensa ou punição podem ser associados a percepções, e mais tarde, em caso de percepções similares, ativar os respectivos sinais de prazer ou desprazer. Quando ocorre essa associação é dito que há um significado P/DP (funcionalmente similar ao significado emocional em sistemas biológicos).

Consciência para Haikonen

No sistema cognitivo de Haikonen o conteúdo da consciência é criado pela percepção. O conhecimento do ambiente é criado pela percepção do ambiente, assim como a autoconsciência é criada pela percepção do corpo e de seus processos, além da percepção do conteúdo mental (introspecção). Não há circuitos especiais ou locações em que os circuitos se tornam conscientes. Tanto no modo "consciente" como "inconsciente", cada circuito opera basicamente da mesma maneira. O conteúdo se torna "consciente" quando vários circuitos operam em uníssono, com a mesma entidade sendo foco de vários módulos (Haikonen, 2003, Cap. 18).

O modelo de Haikonen não deve ser visto como um modelo para o cérebro, mas como

um esboço para uma máquina cognitiva. As questões de consciência estão relacionadas com a conversação interna e imagens mentais que permitem um certo grau de auto-consciência e introspecção ao sistema, o que seria comparável à ideia de consciência de acesso de Block[6]. Haikonen afirma que seu sistema pode ser comparado com o modelo da teoria do *workspace* global[7] de Bernard Baars. Esse paralelo pode ser feito principalmente se observados os pontos de percepção como transmissores globais, os quais podem ser comparados com os "palcos do teatro" de Baars (Haikonen, 2000b). Na interpretação de Haikonen, o palco do teatro é uma memória de trabalho que oferece conversação interna e imagens mentais (Haikonen, 2003, p. 157).

3.3.2 CLARION

A arquitetura cognitiva *CLARION* (*Connectionist Learning with Adaptive Rule Induction ON-line*[8]), tem sido desenvolvida por Ron Sun, professor de ciências cognitivas do Instituto Politécnico de Rensselaer - EUA. As ideias iniciais foram apresentadas em (Sun *et al.* , 1996; Sun, 1997, 1999; Sun *et al.* , 2001) e de forma mais detalhada em (Sun, 2003).

Vários experimentos relacionados com o aprendizado de habilidades foram realizados na arquitetura CLARION. Entre elas, estão tarefas relacionadas ao aprendizado implícito como: tarefas de tempo de reação serial[9] (Sun & Terry, 2002; Sun *et al.* , 2005), aprendizado de gramáticas artificiais (Sallas *et al.* , 2007) e tarefas de controle de processos. Também foram realizados testes com tarefas de aquisição de habilidades cognitivas de alto nível (envolvendo um grande número de processos explícitos) como, por exemplo, Torre de Hanoi (Sun *et al.* , 2005) e aritmética alfabética. Além disso, a arquitetura foi testada na tarefa de navegação em campo minado (Sun, 1997; Sun & Peterson, 1998; Sun *et al.* , 2001), em simulações de motivação e metacognição, e em tarefas de simulação social (Sun & Naveh, 2004, 2007).

A maioria das arquiteturas cognitivas tem focado em modelos *top-down* em que é realizado o aprendizado do conhecimento explícito primeiramente para, então, baseado nele, realizar o aprendizado implícito. A direção *bottom-up* é normalmente negligenciada pela literatura (Sun *et al.* , 2009). Nesse contexto, a arquitetura CLARION é um modelo integrador com duas estruturas de representação. É formado por dois subsistemas principais: *subsistema centrado em ações*[10] (ACS) e *subsistema não centrado em ações*[11]. Ambos os subsistemas são divididos em dois níveis: no nível superior é

[6] ver seção 2.2.1.

[7] ver seção 2.3.8.

[8] CLARION foi implementado em Java e pode ser encontrado em http://www.cogsci.rpi.edu/ rsun/clarion.html.

[9] Em um exemplo desse tipo de tarefa (em inglês chamada de *serial reaction time task*), uma dica visual pode aparecer na tela de um computador, organizado horizontalmente em quatro posições. Cada posição corresponde a um botão no teclado. A cada experimento o participante seleciona o botão referente à posição. Esse tipo de experimento tem sido utilizado por psicólogos para explorar os processos relacionados a uma vasta gama de comportamentos incluindo princípios cognitivos e biológicos do aprendizado e memória (Robertson, 2007).

[10] em inglês *action-centered subsystem*.

[11] em inglês *nonaction-centered subsystem*.

Tabela 3.1: Comparação das duas dimensões da arquitetura *CLARION* (reprodução de (Sun & Franklin, 2007))

Dimensão	**nível inferior**	**nível superior**
Fenômeno cognitivo	aprendizado implícito	aprendizado explícito
	memória implícita	memória explícita
	processamento automático	processamento controlado
	intuição	raciocínio explícito
Fontes de conhecimento	tentativa e erro	fontes externas
	assimilação de conhecimento explícito	extração do nível inferior
Representação	(micro) características distribuídas	unidades conceituais locais
Operação	baseada em similaridade	manipulação explícita de símbolos
Características	mais sensível ao contexto (fuzzy)	mais exato (crisp), precisa
	menos seletiva	mais seletiva
	mais complexa	mais simples

codificado o conhecimento explícito e, no nível inferior, o conhecimento implícito[12].

ACS

O subsistema centrado em ações executa a tomada de decisão baseado no seguinte algoritmo (Sun *et al.* , 2009):

1. Observe o estado atual x
2. Compute, no nível inferior (i.e. a IDN[13] ou Rede de Decisão Implícita), o "valor" de cada uma das possíveis ações (a_i's) no estado atual x : $Q(x, a_1), Q(x, a_2), ..., Q(x, a_n)$. Estocasticamente escolha uma ação.
3. Encontre todas as possíveis ações $(b_1, b_2, ..., b_m)$ no nível superior (i.e. a ARS[14], ou a Armazém de Regras de Ação), baseados na informação do estado atual x (que veio do nível inferior) e nas regras existentes já instauradas no nível superior. Estocasticamente escolha uma ação.
4. Encontre uma ação apropriada a, selecionando estocasticamente baseado na combinação dos valores a_i's (da saída do nível inferior) e dos b_j's (da saída do nível superior).
5. Execute a ação selecionada a, e observe o próximo estado y.

[12] ver figura 3.3 para uma visão geral da arquitetura e a tabela 3.1 para características gerais de cada nível.

[13] em inglês *Implicit Decision Network*.

[14] Do inglês, Action Rule Store

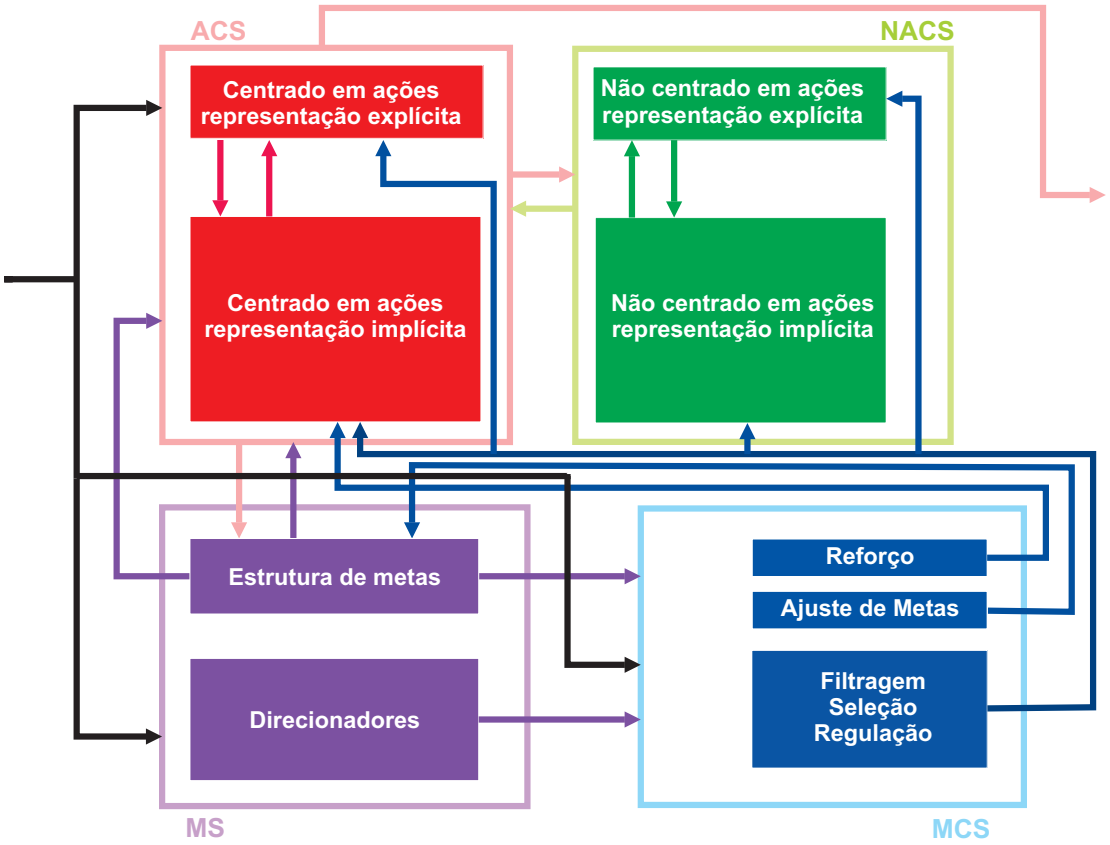


Figura 3.3: Implementação da arquitetura CLARION. ACS indica o subsistema centrado em ações, NACS o subsistema não centrado em ações, MS o subsistema motivacional, MCS o subsistema metacognitivo. O nível superior contém codificação local de conceitos e regras. O nível inferior contém redes conexionistas múltiplas (modulares) para capturar processos inconscientes. As interações dos dois níveis e o fluxo de informações estão indicados pelas setas. Adaptado de (Sun & Franklin, 2007, Figura 7.3)

6. Atualize os valores Q no nível inferior de acordo com *Q-learning* (implementado com uma rede neural com retroalimentação).

7. Atualize o nível superior com um algoritmo de aprendizado apropriado (para construir, refinar, e apagar regras explícitas).

8. Volte ao Passo 1.

Rotinas reativas implícitas são aprendidas no nível inferior do subsistema centrado em ações. Cada *valor-Q* é uma avaliação da "qualidade" de uma ação em um dado estado. $Q(x, a)$ indica quão desejável é a ação a no estado x. Um estado x é representado por um conjunto de pares *(dimensão, valor)*: $(dim_1, val_1), (dim_2, val_2), ..., (dim_n, val_n)$, vindos das entradas sensoriais. No nível inferior cada par é enviado para o nó correspondente na rede. No nível superior cada par é testado nas condições das regras (Sun *et al.* , 2009).

O *valor-Q* pode ser calculado através do algoritmo *Q-Learning* (Watkins, 1989), um algoritmo de aprendizado por reforço que compara os valores das ações sucessivas e, baseado nessa comparação, ajusta a função de avaliação. Assim, ele desenvolve comportamentos sequenciais reativos e fornece ao agente a possibilidade de escolher uma ação a qualquer momento (escolhendo a de maior *valor-Q*, por exemplo)[15] (Sun & Franklin, 2007).

No passo 4 a seleção de probabilidades é determinada por um processo de *"casamento de probabilidades"*. Nesse processo, os dois níveis competem para terem suas ações utilizadas no direcionamento do agente. O desempenho de cada nível é calculado através da *taxa de casamento positivo*, ou seja, a razão de *casamentos positivos*[16] que um determinado componente produziu sobre todos os casamentos. As taxas de sucesso de cada componente são então utilizadas para determinar a seleção de probabilidades dos diversos componentes (Sun *et al.* , 2009).

O nível inferior do subsistema centrado em ações possui algumas pequenas redes neurais que coexistem. Cada uma delas é adaptada para tarefas ou grupos de entrada sensorial específicas, criando uma modularidade, como os grupos neurais de Edelman[17] ou os processadores de Baars[18] (Sun & Franklin, 2007). Nessa teoria, a maior parte do processamento da mente humana é realizada por processadores especializados altamente eficientes. Na arquitetura CLARION alguns desses módulos são embutidos no sistema (como as habilidades inatas nos seres humanos) e outros podem ser obtidos por aprendizado através da interação com o ambiente como realizado em (Sun & Peterson, 1999; Sun & Sessions, 2000; Sun, 2003, p. 18).

O conhecimento explícito do nível superior do subsistema centrado em ações está codificado na forma de regras de ação que podem ser aprendidas das mais diversas maneiras. Em linhas gerais, o agente dinamicamente adquire representações e as modifica conforme necessário, refletindo a natureza dinâmica da habilidade de aprendizado (Sun & Peterson, 1998; Sun *et al.* , 2001).

O aprendizado *bottom-up* é realizado através do algoritmo RER[19] o qual aprende regras utilizando informações do nível inferior. No algoritmo RER, se uma ação decidida no nível inferior for bem sucedida (ou seja, se ela satisfizer um certo critério), então o agente extrai a regra e a adiciona ao conjunto de regras do nível superior e a aplica nas próximas interações com o ambiente por generalização ou especialização das condições da regra (Sun *et al.* , 2009). Na direção *top-down*, as regras existentes no nível superior - as quais são chamadas de regras fixas, ou FR [20] - guiam o aprendizado do nível inferior. Inicialmente, o agente se baseia nas FRs para a tomada de decisão das ações. O nível inferior adquire cada vez mais conhecimento, por meio da observação das ações direcionadas pelas regras. Devido a isso o agente pode dar mais credibilidade às ações do nível inferior, o que habilitaria o aprendizado *top-down* (Sun *et al.* , 2009).

[15] Para detalhes da implementação e justificativas cognitivas veja Sun *et al.* (2001).

[16] Um critério determinando o que é um *casamento positivo* deve ser definido. Para mais detalhes ver (Sun, 2003).

[17] ver seção 2.3.3.

[18] ver seção 2.3.8.

[19] em inglês Rule-Extraction-Refinement.

[20] em inglês *fixed rules*.

NACS

Enquanto o ACS envolve o que tradicionalmente é chamado de conhecimento procedimental, o subsistema não centrado em ações (NACS) inclui conhecimento (implícito e explícito) geral sobre o ambiente , o que é comumente chamado de "memória semântica" (Sun *et al.* , 2009). Esse subsistema é controlado através das ações do ACS e é possível tanto recuperar informações como executar diversas inferências.

Sun *et al.* (2009) propõe que o sentido de implícito/explícito de um conhecimento está relacionada à facilidade de acesso e de uso, independente de processos ou mecanismos de acesso, seguindo a linha apresentada em (Kirsh, 1990):

> a explicitação [da representação] realmente está relacionada a quão rapidamente a informação pode ser acessada (...). Ela está mais associada ao que é apresentado em um sentido de processo, do que ao que é apresentado em um sentido estrutural.

De um lado, a natureza inacessível dos conhecimentos implícitos é capturada por representações subsimbólicas, distribuídas através de redes neurais com retropropagação, que mapeiam entradas e saídas (Gaglio, 2007; Sun *et al.* , 2009). Em geral, as unidades de representação em uma representação distribuída são capazes de realizar tarefas mas são subsimbólicas e sem sentido individualmente (Sun, 1995, 1999). Essa representação distribuída é menos acessível estando de acordo com a (relativa) inacessibilidade do conhecimento implícito (Sun, 1999; Sun *et al.* , 2009).

Por outro lado, o conhecimento explícito é obtido por uma modelagem computacional de representações locais ou simbólicas pelas quais cada unidade de representação é mais facilmente interpretada e há um significado conceitual claro. Essa característica lembra os conhecimentos explícitos por serem de mais fácil acesso e manipulação. Assim, a rede de alto nível é formada por *"chunks"* os quais são especificados por valores (características) dimensionais. Esses *chunks* têm o formato: $chunkid_i : (dim_{i1}, val_{i1})(dim_{i2}, val_{i2})...(dim_{in}, val_{in})$, em que *dim* denota a dimensão do estado/saída e *val* especifica o valor correspondente. Os *chunks* são armazenados em uma rede chamada memória de conhecimento geral (GKS[21]), em que cada nó é um *chunk*. Esses nós são conectados aos nós de características correspondentes no nível inferior. Além disso, há ligações entre *chunks* que codificam associações entre pares de *chunks* (Sun *et al.* , 2009).

MS

O subsistema motivacional (MS) é formado por *"drives"*[22] e suas interações, que levam às ações. Basicamente, o agente escolhe as ações que maximizam seu ganho, recompensa, reforço, ou retornos. Esse subsistema influencia o subsistema centrado em ações por fornecer um contexto em que estão definidos metas e o reforço do subsistema centrado em ações. Desse modo o MS explica porque o agente faz o que ele faz (Sun, 2007).

[21] em inglês, *General Knowledge Store*.

[22] Poderia ser traduzido por direcionadores, objetivos ou metas. Entretanto, será mantido a palavra original em inglês pois essas traduções não se demonstrarem perfeitamente adequadas.

Em *CLARION* há um sistema bipartido de representação motivacional. As metas explícitas[23] de um agente, por exemplo *"encontre comida"*, podem ser geradas através de estados de *drives* internos, como *"estar com fome"*[24]. Além dos *drives* de baixo nível, os quais estão relacionados a necessidades fisiológicas, há os *drives* de mais alto nível. Alguns deles são primários, por já estarem embutidos no agente e serem relativamente inalteráveis (Sun, 2007).

MCS

O subsistema metacognitivo (MCS) trabalha em conjunto com o MS. O MCS monitora, controla e regula os processos cognitivos com a finalidade de melhorar o desempenho cognitivo (Sun, 2003).

A regulação e o controle acontecem na forma de ajuste de metas e parâmetros essenciais para o ACS e o NACS e na interrupção ou mudança nos processos em andamento. O controle e a regulação também podem ocorrer através da definição das funções de reforço do ACS. Essas ações são realizadas com base nos estados dos direcionadores do MS (Sun, 2007).

Assim como os outros módulos, o MS também é subdividido em dois níveis: o explícito (nível superior) e o implícito (nível inferior).

Consciência na CLARION

CLARION é um modelo de dois níveis: superior, que possui representações simbólicas e inferior, o qual possui representação distribuídas. Há ainda um tipo de método de aprendizado para cada nível. No desenvolvimento do modelo, quatro hipóteses foram consideradas (Sun & Franklin, 2007):

1. acessibilidade direta dos processos conscientes, figurada pelo conhecimento consciente das representações simbólicas;

2. inacessibilidade direta dos processos inconscientes, representada pelo conhecimento inconsciente inerente à representação subsimbólica das redes neurais de retropropagação utilizadas;

3. ligação entre os conceitos simbólicos (nível superior) às características distribuídas (nível inferior): quando o conceito simbólico está ativado, suas representações distribuídas também são ativadas (Sun, 1995);

4. ligações entre as representações distribuídas e os conceitos simbólicos: sob determinadas circunstâncias, uma vez que algumas ou a maioria das características distribuídas de um conceito são ativadas, os conceitos simbólicos relacionados podem se ativar para "cobrir" essas características.

[23] as quais estão ligadas ao trabalho do subsistema centrado em ações.
[24] para detalhes veja (Sun, 2003).

Sun (1999) defende que a sinergia entre o aprendizado implícito (inconsciente) e explícito (consciente) está ligada à consciência. Nesse sentido, a característica da consciência de poder vetar ou contrabalancear decisões do inconsciente pode ser vantajosa para a performance geral do sistema. Em relação à inconsciência, na arquitetura CLARION os processos de memória implícita, aprendizado implícito, percepção inconsciente e automatização trazem à tona informações inconscientes que são registradas no nível inferior. Isso permite o uso dessas informações quando necessário.

3.4 Outros trabalhos

3.4.1 Robôs Auto-Conscientes de Takeno

Junichi Takeno, professor do departamento de Ciências da Computação da Universidade de Meiji no Japão, e seu grupo de pesquisa têm desenvolvido robôs capazes de diferenciar quando estão vendo a si mesmos em um espelho ou quando estão observando um robô idêntico (Inaba & Takeno, 2003; Takeno *et al.* , 2005; Suzuki *et al.* , 2005).

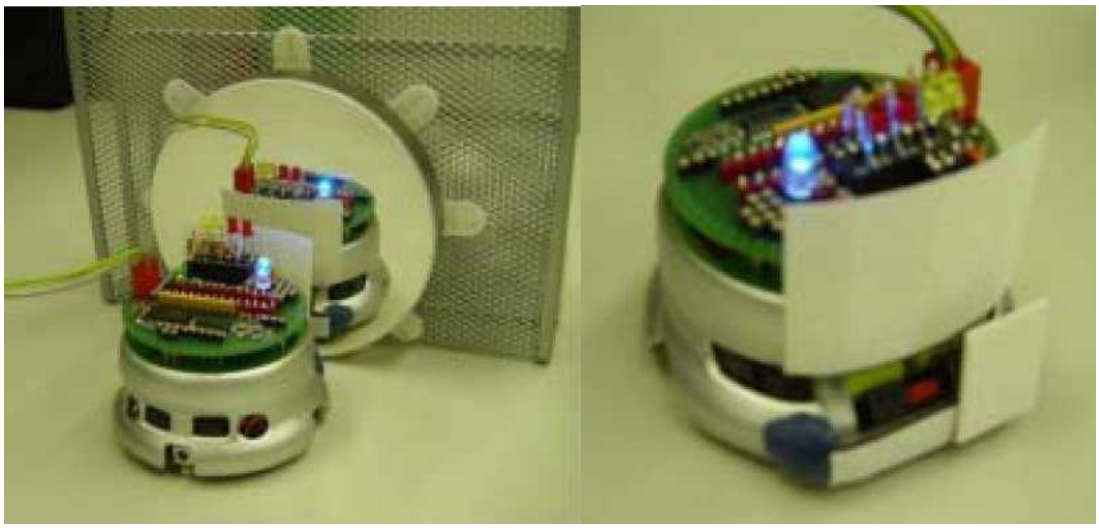


Figura 3.4: Robô auto-consciente de Takeno. Fonte: Instituto de Pesquisas em Ciências Heurísticas, Japão

Suzuki *et al.* (2005) defende que a imitação é um ato de consistência entre cognição e comportamento. Por um lado, é necessária a separação entre o indivíduo, os demais e o ambiente. Por outro lado, a imitação significa aprender o comportamento dos outros e, ao mesmo tempo, transferi-lo para si, entendendo e integrando as condições internas e externas. Assim, define que a consciência vem da consistência de cognição e comportamento.

Para resolver a questão do paralelismo entre o processamento de informação e a execução de comportamentos, Takeno e seu grupo desenvolveram uma rede neural recursiva que é comum aos dois processos. Takeno criou um sistema de sinalização para mostrar o modo de funcionamento do robô. Alguns LEDs foram conectados diretamente à rede

neural e acendem conforme o tipo de comportamento. Quando o LED vermelho está aceso, significa que o robô está executando seu próprio comportamento. Se ele percebe outro robô executando algo, o LED verde é aceso. Por último, se o LED azul acender, isso indica que o robo esta observando um outro robô executar uma ação e buscando imitá-lo.

Nos primeiros resultados, o comportamento de imitação possui uma taxa de coincidência de 70% em frente a um espelho e de 60% em frente a um robô similar, entretanto esse número pode ser melhorado realizando o tratamento de fatores de incerteza inerentes ao experimento (Takeno *et al.* , 2005).

3.4.2 Cicerobot

Cicerobot (Liotta *et al.* , 2005; Chella & Macaluso, 2006; Chella, 2007) é um robô criado pelo Robotics Lab[25], grupo de pesquisa liderado por Antonio Chella, professor e pesquisador de robótica da Universidade de Palermo na Itália. A arquitetura foi montada sobre um robô RWI B21 equipado com câmera, laser para telemetria e um sonar (Figura 3.5). Cicerobot foi desenvolvido para funcionar como um guia do Museu Arqueológico de Agrigento, não somente em *tours* pré-programados, mas também em trajetos gerados através da interação com o público (Chella *et al.* , 2005). Essa tarefa pode ser considerada um caso de estudo relevante por envolver percepção, autopercepção, planejamento e interação homem-máquina (Burgard *et al.* , 1999).

Figura 3.5: Cicerobot. Fonte: Robotics Lab

A arquitetura cognitiva do robô é baseada em um simulador 3D interno que é atualizado conforme a navegação. Todos os comandos dos atuadores são também enviados para o simulador o qual calcula as posições e imagens da câmera. Após executado o

[25] http://roboticslab.dinfo.unipa.it

movimento, Cicerobot compara o estado esperado com o real. A simulação 3D também é utilizada para planejar ações por meio do estudo de diversos cenários e pode ser comparada com a imaginação humana (Chella & Macaluso, 2006).

3.4.3 CRONOS

CRONOS[26] é um projeto para a criação de um robô consciente coordenado pelo professor Owen Holland, do departamento de sistemas eletrônicos, e de computação da Universidade de Essex e pelo professor Tom Troscianko, do departamento de psicologia experimental da Universidade de Bristol na Inglaterra. O projeto consiste em um robô (*CRONOS*) baseado em um sistema muscular e esquelético humano (Holland & Knight, 2006) e no *SIMNOS*, um software de simulação em tempo real desse robô e seu ambiente (ver figura 3.6), além de um sistema visual de inspiração biológica e um simulador neural chamado *SpikeStream* [27].

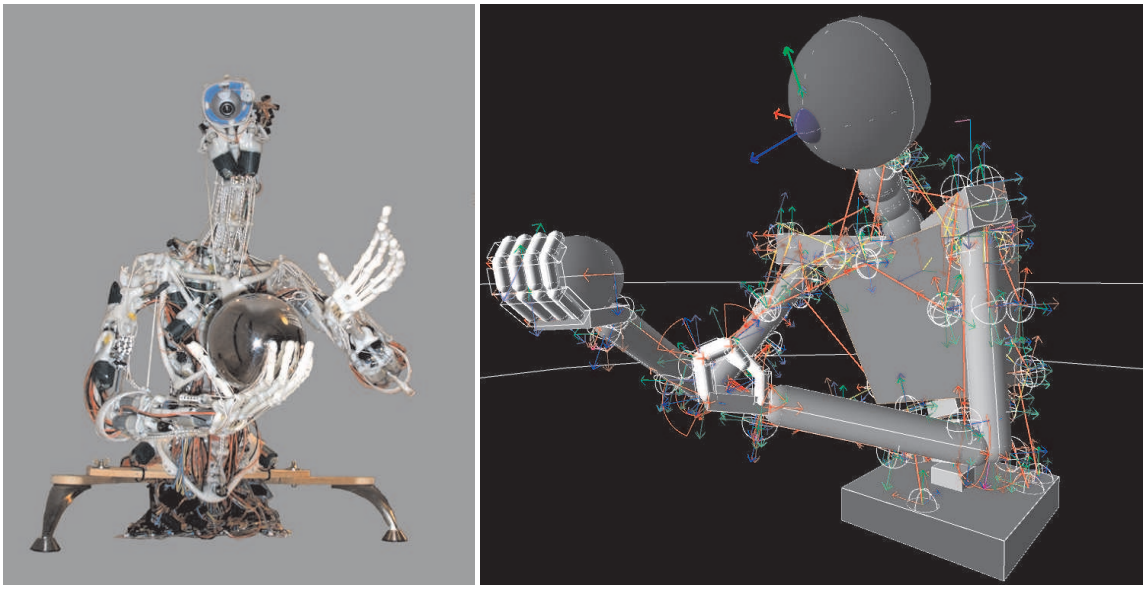


Figura 3.6: Projeto CRONOS. À esquerda o robô CRONOS e, à direita, SIMNOS, o software de simulação. Fonte: Site CRONOS

Nesse projeto, uma abordagem é a realizada por Holland, que afirma que modelos internos têm um papel fundamental nos estados cognitivos conscientes e podem ser a causa da consciência humana, ou estarem com ela correlacionados. Holland *et al.* (2007) se concentraram nos modelos internos que incluem o corpo do agente e o seu relacionamento com o ambiente, uma discussão que tem conexão com o modelo de *self* de Damásio (ver seção 2.3.2). O robô utiliza o simulador SIMNOS como modelo interno de si mesmo, o qual é constantemente atualizado e empregado como modo de selecionar as ações antes de executá-las (Gamez, 2008b).

[26] http://www.cronosproject.net
[27] http://spikestream.sourceforge.net

Gamez (2008a) abordou a consciência de máquina nesse projeto de uma outra maneira. Utilizando a SpikeStream, desenvolveu uma rede neural para controlar os movimentos dos olhos de SIMNOS e CRONOS. Essa rede, quando ativa, gera movimentos dos olhos para diferentes partes do campo visual do robô e realiza a associação entre a posição dos olhos e os estímulos visuais. Além disso, a rede possui um sistema emocional que troca o modo de operação para "imaginação" quando um objeto "negativo" é encontrado. Isso inibe as entradas sensoriais e saídas motoras enquanto a rede explora os padrões sensoriais, até que ela encontre um estímulo positivo no sistema emocional. Tal processo desabilita o sistema de inibição e o olho é movido em direção ao objeto selecionado.

3.5 Críticas ao estudo de consciência artificial

Não é incomum, em um período pré-paradigmático pelo qual passa a pesquisa em consciência artificial, que essa área de pesquisa seja alvo de contestações de forma mais contundente. A descrença dos cientistas de computação não deve ser olhada com estranheza, uma vez que a própria consciência humana passou anos tendo a sua existência negada ou as pesquisas relegadas a um segundo plano. Nesse sentido, William James, um psicólogo e filósofo americano do final do século 19 afirmou:

> Nos últimos vinte anos eu não acreditei na "consciência" como uma entidade; nos últimos sete ou oito anos eu sugeri a sua *não* existência aos meus alunos, e tentei dar a eles um equivalente pragmático em realidades da experiência. Parece-me que é chegada a hora disso ser universal e abertamente descartado (James, 1904).

Além disso, diversas outras linhas de pesquisas em inteligência computacional sofreram severas críticas em seu início, como a computação evolutiva, as redes neurais e os sistemas *fuzzy*, e hoje são empregadas com sucesso em várias soluções computacionais na indústria.

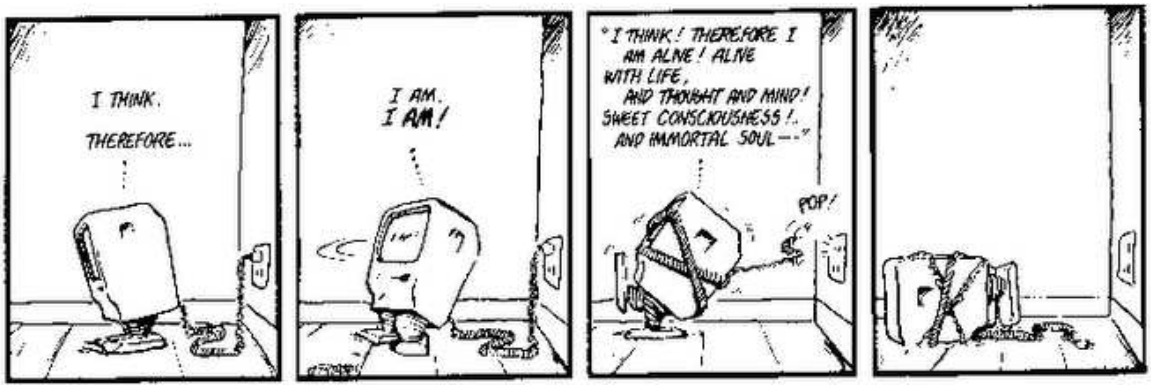


Figura 3.7: Crítica às máquinas conscientes

Haikonen (2003, p. 162) agrupa algumas objeções ao estudo de consciência artificial. Várias delas são baseados na visão dualista[28] de que uma máquina não poderia produzir uma mente (não material). Alguns críticos buscam criar um paralelo entre o homem e a máquina, afirmando que como os seres humanos não são máquinas ou que as máquinas não são vivas, elas não podem ser conscientes. Outros se baseiam em uma suposta posição privilegiada do homem, o qual seria superior às máquinas e aos animais, sendo que a consciência seria o que distinguiria a humanidade dos demais seres vivos. Por fim, há os que se preocupam com questões sociais, e defendem que máquinas conscientes poderiam por em risco a cultura, a ética e a moral.

Nesse contexto, a maioria dos críticos aponta a dificuldade de resolver o problema difícil da consciência[29]. Mesmo tendo em vista que o *problema fácil* não é tão fácil, pelo que se pode encontrar na literatura, é possível afirmar que há um direcionamento na resolução de vários de seus desafios. Já no caso do *problemas difíceis*, apesar de algumas teorias esboçarem a sua solução, os cientistas não apresentam uma conclusão concreta para resolvê-lo. Sem o entendimento de como é produzida a consciência humana, parece sem sentido produzir um robô com consciência fenomenal. Gamez (2008a) elenca algumas razões de porque, mesmo assim, esses estudos são válidos:

- as máquinas com modelos conscientes podem auxiliar o entendimento da própria consciência humana;
- a dificuldade de se dizer se uma máquina é consciente ou não abre espaço aos cientistas para afirmarem que uma máquina é consciente, mesmo não sabendo afirmar se é um caso de solução do *problema difícil*;
- há pesquisas que indicam a possibilidade de se criar condições para emergir consciência em um sistema, mesmo sem o entendimento das causas dos estados fenomenais (se é que de fato eles existem!).

Outro argumento utilizado remete ao quarto chinês de (Searle, 1980). O quarto chinês consiste em uma pessoa que recebe caracteres em chinês, processa de acordo com um conjunto de regras em sua própria língua e passa os resultados em chinês sem realmente compreender o idioma. Assim, seguindo essa lógica, seria possível modelar uma mente em um robô sem realmente ele possuir estados fenomenais.

Por último, Penrose, utilizando o teorema da incompletude matemática de Gödel, defende que a consciência humana não pode ser nem mesmo simulada em uma máquina de Turing clássica como um computador (Penrose, 1989, 1994). Ainda que Penrose esteja correto, há a possibilidade de se criar um computador quântico com os mecanismos físicos relatados por Penrose para a consciência humana (Gamez, 2008b).

3.6 Conclusão

Esse capítulo sintetizou os principais trabalhos no campo de consciência artificial e as críticas no campo de estudo. Como o campo de consciência artificial é ainda pré-

[28] ver seção 2.2.2.
[29] ver seção 2.2.4.

paradigmático, diversos conceitos e abordagens aparecem nos estudos apresentados.

O próximo capítulo encerra o estudo teórico das arquiteturas cognitivas apresentando a arquitetura Baars-Franklin e detalhando os seus mecanismos.

Capítulo 4

Arquitetura Baars-Franklin e seus Mecanismos

> A razão pela qual a maioria das pessoas não atinge os objetivos é porque não os conseguem definir ou não os consideram atingíveis. Os vencedores conseguem sempre dizer-lhe para onde estão indo, o que pretendem fazer ao longo do caminho e quem irá compartilhar a aventura com eles.
>
> Denis Watley

4.1 Introdução

Esse capítulo encerra o estudo teórico das arquiteturas cognitivas apresentado na seção 3.3 e traz de forma mais detalhada a arquitetura *Baars-Franklin*[1] desenvolvida pelo grupo de Stan Franklin, da Universidade de Memphis, EUA.

A arquitetura Baars-Franklin nasceu com o objetivo de ser uma arquitetura cognitiva para o modelo de consciência da teoria do workspace global de Baars e tem sido utilizada em várias aplicações. Os principais exemplos de uso dessa arquitetura são os agentes de software: CMattie (Franklin & Graesser, 1999) e IDA (Franklin, 2005), desenvolvidos pelo grupo de Franklin, e CTS (Dubois, 2007a), implementado por Daniel Dubois da Universidade de Quebec, os quais são descritos na seção 4.2.

A seguir, na seção 4.3 é mostrada uma visão geral dos diversos componentes da arquitetura e, em seguida, é apresentado o funcionamento dos *codelets* (seção 4.4) e dos

[1]Normalmente a arquitetura computacional proposta por Franklin é chamada de *tecnologia IDA*, havendo uma sobreposição de nomes com "o agente IDA". Para diferenciá-los, nesse trabalho a arquitetura cognitiva computacional proposta por Franklin e baseada na teoria do *workspace* global de Baars será chamada de *arquitetura Baars-Franklin*.

módulos de percepção (seção 4.5), memória associativa (seção 4.6) e memória episódica (seção 4.7). O mecanismo de consciência e a rede de comportamentos, os dois módulos principais da arquitetura, são apresentados, respectivamente, nas seções 4.8 e 4.9. Por fim, são mostradas na seção 4.10 as interações entre os diversos componentes da arquitetura através da descrição do ciclo de operação da arquitetura, chamado de *ciclo cognitivo*.

4.2 Contexto histórico

IDA - Intelligent Distribution Agent (Franklin *et al.* , 1998; Franklin, 2003, 2005) é um agente de software "consciente"[2] que foi desenvolvido pelo *Cognitive Computing Research Group*[3], grupo liderado por Stanley Franklin, matemático e pesquisador das ciências cognitivas e de computação da Universidade de Memphis. Esse agente foi desenvolvido para elaborar, de maneira autônoma, a distribuição de trabalhos aos marinheiros da marinha dos Estados Unidos, tarefa realizada manualmente por 280 pessoas (McCauley & Franklin, 2002).

IDA é uma evolução de uma série de estudos realizados desde 1996 por mais de 25 pesquisadores. O agente *VMattie*[4] (Song, 1998) foi o precursor de várias ideias do grupo e não trazia características "conscientes" como a arquitetura atual. O primeiro agente "consciente" surgiu com a *CMattie*[5] (Ramamurthy *et al.* , 1998; Franklin & Graesser, 1999; Anwar & Franklin, 2003), agente irmão de IDA e fonte de inspiração para o desenvolvimento de *ConAg* (Bogner, 1999; Bogner *et al.* , 1999), um framework para implementação de "consciência" em agentes de softwares.

Tanto *VMattie* quanto *CMattie* tinham o papel de coordenar as informações dos seminários semanais do departamento de matemática da Universidade de Memphis. O objetivo dos agentes era recolher as informações dos humanos sobre seminários e eventos como colóquios, defesas, etc; usar essas informações e compor o anúncio dos eventos da semana seguinte, novamente utilizando para isso um sistema de mensagens eletrônicas (Bogner, 1999).

Mais tarde, Daniel Dubois desenvolveu o agente *CTS - Conscious Tutoring System* (Dubois *et al.* , 2006; Dubois, 2007a,b), um tutor consciente baseado na arquitetura Baars-Franklin. Esse agente foi desenvolvido para trabalhar em um domínio completamente diferente: auxiliar o treinamento de astronautas na aprendizagem do uso do braço mecânico *Canadarm2*, da ISS - Estação Espacial Internacional (figura 4.1).

Atualmente, o grupo de pesquisa de Stan Franklin tem buscado aprimorar as habilidades de aprendizado do agente, criando a *LIDA* - Learning IDA (Franklin & Ferkin, 2006; Ramamurthy *et al.* , 2006; Friendlander & Franklin, 2008).

[2] Para Stan Franklin e seu grupo de pesquisa, um agente de software *"consciente"* nada mais é do que um agente de software que integra vários mecanismos de inteligência artificial, a fim de implementar a *Teoria do Workspace Global* de Baars (Bogner, 1999, p. 49). Franklin (2003) não vê argumentos convincentes para afirmar que *IDA* é fenomenalmente consciente.

[3] Página do grupo: http://ccrg.cs.memphis.edu/.

[4] Virtual Mattie.

[5] Conscious Mattie.

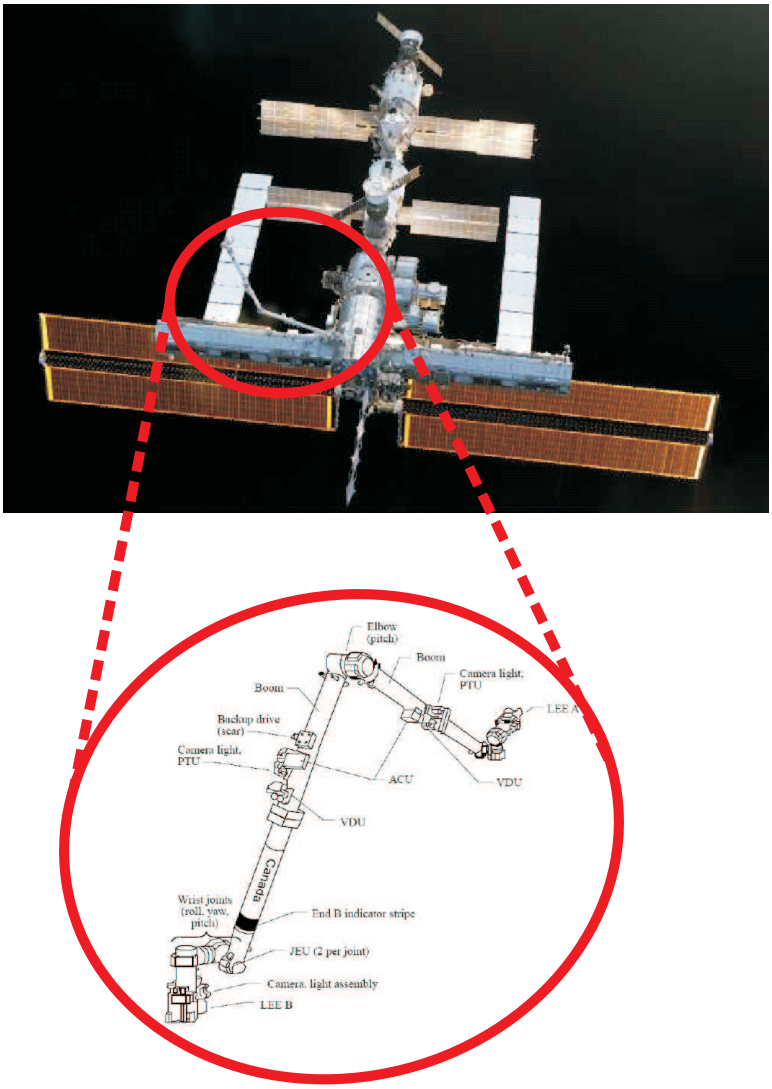


Figura 4.1: Braço mecânico Canadarm2. Adaptado de (Dubois, 2007a)

4.3 Visão Geral

A arquitetura Baars-Franklin realiza uma série de processos cognitivos como percepção, seleção de ações, aprendizado, emoções, "consciência", automatização, resolução de problemas e metacognição. Para isso, a arquitetura se baseia em várias tecnologias anteriores, incluindo: rede de comportamentos (Maes, 1989), arquitetura *Copycat* (Hofstadter, 1994), memória esparsa distribuída (Kanerva, 1991) e teoria do Pandemônio (Selfridge, 1958; Jackson, 1987). A arquitetura é dividida em diversos módulos (ver figura 4.2) que são apresentados a seguir.

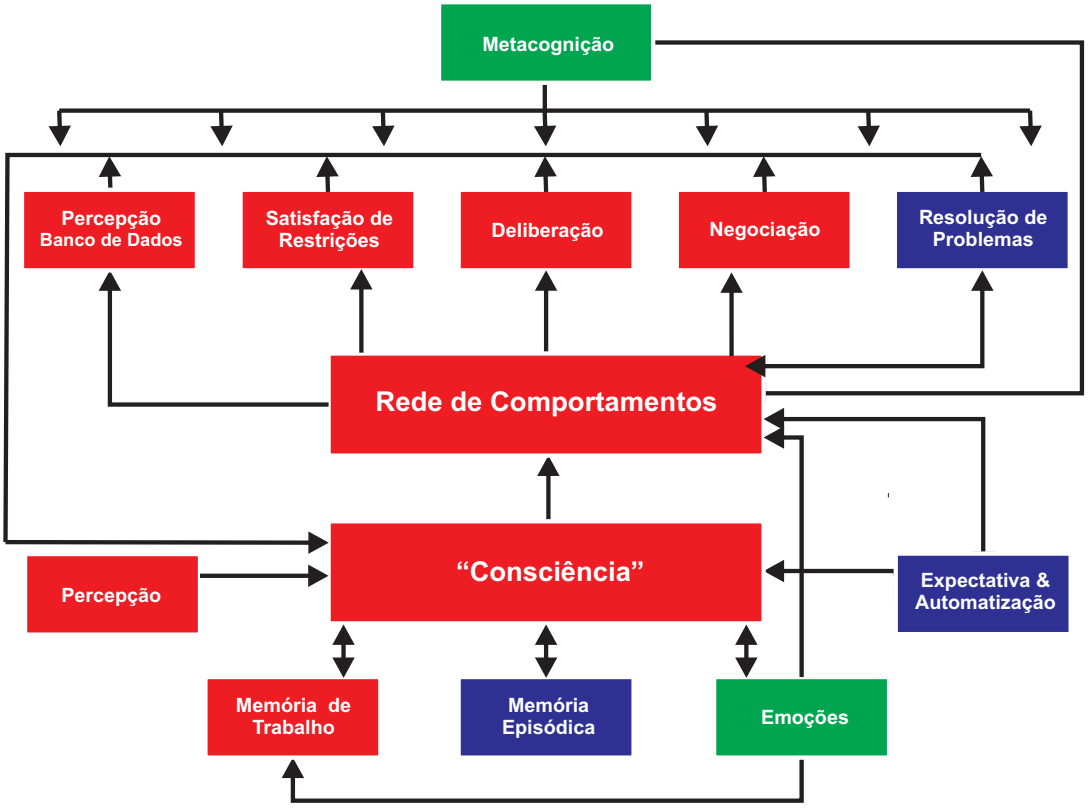


Figura 4.2: Arquitetura *Baars-Franklin* em *IDA*. Em vermelho estão os módulos implementados, em azul os módulos em desenvolvimento e, em verde, a parte projetada da arquitetura. Adaptado de (Franklin & Jr, 2006)

4.4 Codelets

Codelets são processadores especialistas, agentes simples, relativamente independentes e de propósito específico. O nome "codelet" vem da terminologia empregada em (Hofstadter, 1994) e corresponde aos processadores especializados que, segundo a teoria do *workspace* global, são a base da cognição humana. Cada codelet é implementado como *thread* independente (Bogner, 1999). Apesar de ser especializado na realização de tarefas simples, em geral, um codelet trabalha associado a outros codelets. Dessa maneira, os codelets formam *coalizões*, a fim de construir comportamentos de alto nível. Segundo (Negatu, 2006, p. 33), os codelets também podem se comportar como *demons*[6] aguardando o aparecimento de uma determinada condição no ambiente para daí executar a sua tarefa, como no *Pandemônio* de (Selfridge, 1958; Jackson, 1987). Além disso, os codelets dão à arquitetura Baars-Franklin as características de um sistema multi-agente em um único sistema cognitivo como em *Sociedade da Mente* de (Minsky, 1986).

Os codelets armazenam o *conhecimento procedimental* através das atividades primitivas neles codificadas. A relevância desse conhecimento em determinado contexto

[6] Referente aos *demons* de (Selfridge, 1958).

é medida através de um *nível de ativação* do codelet. Codelets se associam a outros codelets, o que pode se transformar em uma coalizão[7], dependendo do *nível de ativação* da *associação*. Associações com alto nível de ativação também servem para criar uma ponte para interações de baixo nível entre os codelets, além de poderem ser vistas como uma rede de codelets simultaneamente ativos da teoria do Pandemônio (Bogner, 1999; Negatu, 2006).

Os codelets, podem ser divididos em *codelets de percepção, de atenção, de informação, de comportamento e de expectativa*, dependendo do módulo que fazem parte. Essa taxonomia não é fixa, podendo ser criados novos tipos de codelet dependendo da necessidade da aplicação. Esses tipos de codelet são apresentadas a seguir.

4.5 Percepção

O módulo de percepção presente na arquitetura Baars-Franklin, utilizado também na construção de IDA, é muito similar ao de VMattie, exposto em (Zhang *et al.* , 1998). Esse módulo realiza o processamento de *strings*, que podem vir por email ou dos registros de um banco de dados. A simplicidade do domínio da aplicação permitiu a utilização de análise probabilística (ao invés do processamento por *parsers* simbólicos convencionais) durante o processo de entendimento da linguagem natural. IDA utiliza um mecanismo de casamento dos tipos de mensagens recebidas com modelos pré-estabelecidos (Franklin, 2003). Esse mecanismo é parte da arquitetura *Copycat* de (Hofstadter, 1994).

O *conhecimento semântico/conceitual* toma a forma de uma rede semântica com níveis de ativação, chamada de *Slipnet*, nome inspirado na arquitetura *Copycat*. Os nós da rede são os símbolos de percepção do agente, no sentido de (Barsalou, 1999), em que, no caso de *IDA*, cada símbolo representa um tipo de mensagem conhecido com uma série de características esperadas. Além disso, diversos *codelets de percepção* buscam, nas mensagens que chegam, por palavras ou frases conhecidas. Quando encontram, os nós apropriados da *Slipnet* são ativados e essa ativação é transmitida pela rede até que ocorra a estabilização. Um ou vários nós-modelo são selecionados através do grau de ativação e os modelos são preenchidos pelos codelets com as palavras/frases selecionadas das mensagens (ver figura 4.3). Isso é o que é chamado de *entendimento da mensagem*, na percepção passiva (Negatu, 2006).

O módulo de percepção foi projetado sob a premissa de que o entendimento da linguagem humana depende da combinação *bottom-up/top-down* de passagem da ativação através da rede conceitual hierárquica, em que os conceitos mais abstratos são colocados no meio. Um nó da Slipnet representa o *conhecimento declarativo* e o seu nível de ativação representa o grau de relevância para a situação corrente do agente. No mecanismo herdado da arquitetura Copycat, conceitos de alto nível emergem através da dinâmica interna de transferência de ativação na Slipnet e das ações dos *codelets de percepção*, que, quando habilitados, ativam os nós-modelo da rede (ver figura 4.3).

[7]Esse mecanismo é coerente com a ideia de coalizão apresentada por Crick e Koch, seção 2.3.4. Outros detalhes são apresentados a seguir na apresentação do módulo de consciência (seção 4.8).

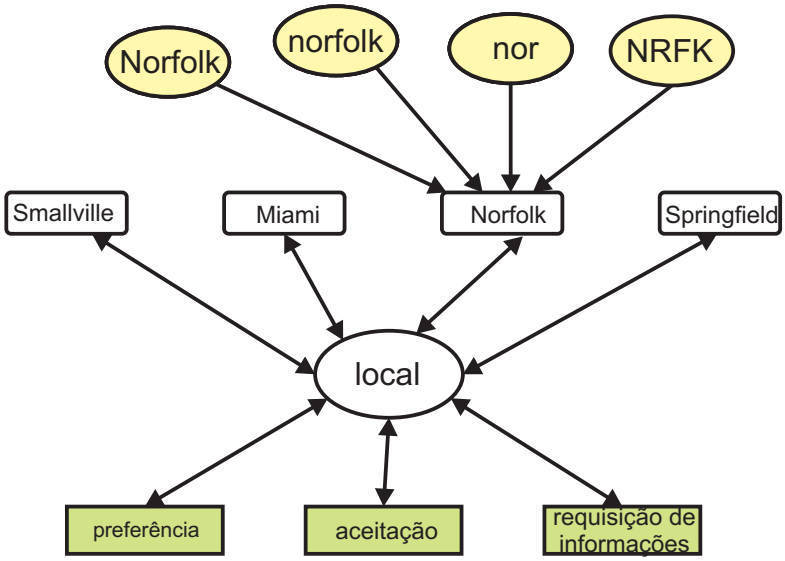


Figura 4.3: Parte da rede Slipnet. Codelets buscam palavras conhecidas, como nas elipses em amarelo. Ao encontrá-las, passam ativação para os nós-modelos (retângulos de cantos arredondados). Esses por sua vez ativam mensagens padronizadas (retângulos verdes).

4.6 Memória Associativa

A arquitetura Baars-Franklin utiliza uma memória esparsamente distribuída (SDM) como memória associativa de longo prazo (Kanerva, 1988, 1991)[8,9]. Uma SDM é uma memória de conteúdo endereçável, ou seja, parte do seu conteúdo é utilizada como entrada na recuperação de um ítem da memória, sem ter que conhecer o endereço do ítem. Devido a essa característica, ela é ideal para o uso como memória associativa de longo prazo (Anwar & Franklin, 2003).

A SDM tem seu nome devido à alocação esparsa dos locais de armazenamento em um grande espaço de endereçamento binário e da natureza distribuída do modo de armazenamento e recuperação de informações. O espaço binário tem 2^n possíveis locais no espaço semântico em que n é tanto o tamanho da palavra (vetores de zeros e uns), quanto a dimensão do endereço do espaço. A dimensão do espaço determina a riqueza de cada palavra. Como n é grande, em termos práticos somente uma pequena porção desse espaço realmente é implementada em um número limitado de locações reais, chamados de *hard locations* (Anwar *et al.* , 1999). O número dessas locações determina a capacidade da memória. As características são codificadas como um ou mais bits e grupos de características são concatenados em uma palavra.

Durante a *escrita* de uma palavra na memória, a cópia dessa palavra é realizada em todas as *hard locations* suficientemente perto da entrada[10]. Na leitura, uma entrada

[8] Para uma discussão sobre os motivos da escolha dessa memória ver (Franklin, 1995, Cap. 13).

[9] Vários outros tipos de memória estão implicitamente adicionados à arquitetura. Para uma discussão desse tema ver (Franklin *et al.* , 2005).

[10] As distâncias são calculados por *distância de Hamming*

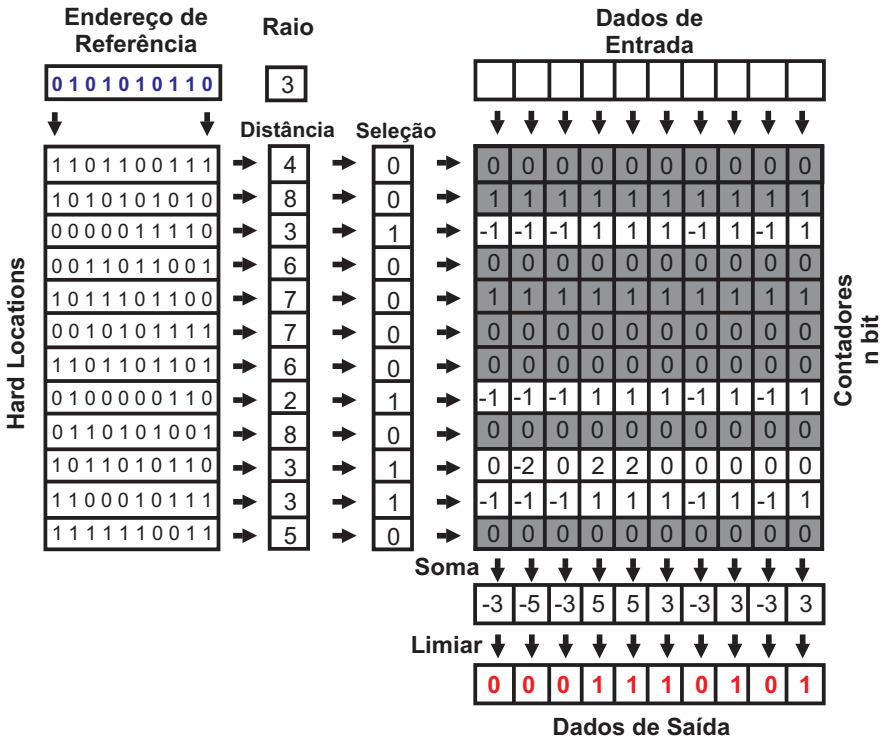


Figura 4.4: Operações de leitura e escrita da SDM. Os *hard locations* ativos são os que possuem distância de hamming menor ou igual a 3. Na operação de leitura, o valor dos contadores associados a esses hard locations ativos são somados. Após verificado o limiar para transformar novamente em saída binária obtém-se os dados de saída. Na operação de escrita, os dados de entrada são somados aos contadores. Adaptado de (Rogers, 1988, Fig. 7).

suficientemente perto pode alcançar vários *hard locations* e o resultado final do processo é um combinado deles (ver figura 4.4). Dependendo da entrada e da informação previamente armazenada na memória, entre outros fatores, a leitura pode não ser bem sucedida. Quando a convergência acontece, a leitura é relativamente parecida com a entrada. Se uma divergência ocorre, a palavra lida não é similar à entrada (Negatu, 2006).

4.7 Memória Episódica Transiente

A memória episódica é a memória que armazena eventos e as suas características de "onde" e "quando". O aprendizado episódico extrai eventos e história das experiências enquanto que o aprendizado semântico extrai fatos do contexto (Tulving, 2002; Nuxoll, 2007).

Nos seres humanos a memória é endereçada por conteúdo, associativa e transiente, com duração da ordem de horas (Conway, 2001). Por exemplo, é possível lembrar os eventos do dia corrente com um certo grau de detalhe, qual o prato do almoço, o que foi

discutido com os amigos durante o coffee-break, etc. Mas esses detalhes são esquecidos conforme o passar do tempo.

Nesse contexto, IDA utiliza uma versão modificada da SDM que possui uma taxa de decaimento em horas (Ramamurthy *et al.* , 2004). Além disso, a SDM modificada utiliza um espaço de memória ternário ("0", "1", e *"don't care"*). Isso possibilita que características desconhecidas sejam representadas. A versão modificada também permite que haja detectores de erros. Dado um evento, a memória episódica deve ser capaz tanto de responder a entradas parciais, como de distinguir um evento memorizado de eventos similares (Ramamurthy *et al.* , 2004).

4.8 Mecanismo de Consciência

Bogner (1999) apresenta o mecanismo de "consciência" e seus principais componentes: *gerenciador de coalizões*, *gerenciador de foco de luz*, *gerenciador de broadcast*, além de uma coleção de codelets de atenção dentre outros, organizados como na teoria do Pandemônio. Esse módulo pode ser compreendido como uma implementação de alguns aspectos do teatro de Baars (Baars, 1997), discutido na seção 2.3.8. A seguir, serão apresentados cada um dos elementos desse mecanismo.

4.8.1 Arena Desportiva

Jackson (1987) faz uma analogia com uma *arena desportiva*, que é formada pela *arquibancada* e pelo *campo de jogo* onde são realizadas as atividades. Nas arquibancadas, estão os codelets inativos, mas que constantemente procuram uma condição relevante para que entrem em campo e se tornem ativos. Quando os codelets se tornam ativos, eles executam as suas tarefas.

4.8.2 Codelets de Atenção

Os *codelets de atenção* reconhecem situações novas ou problemáticas e a sua tarefa é trazer essas informações para a "consciência". Cada codelet de atenção procura por uma determinada situação que pode necessitar de uma intervenção "consciente". Encontrada a situação, o codelet de atenção apropriado se junta a codelets de informação que descrevem a situação. Esse conjunto de codelets forma uma coalizão e os codelets de atenção aumentam o nível de ativação, dependendo da adequação da coalizão à situação esperada. Assim, essa coalizão tem maiores chances de competir pela "consciência"(Negatu, 2006).

4.8.3 Gerenciador de Coalizões

A formação e o monitoramento de coalizões de codelets são feitos pelo *gerenciador de coalizões*. Inicialmente, um codelet ativo não possui associações e pertence a sua própria coalizão. Seguindo a teoria do Pandemônio, os codelets ativos constroem associações com outros codelets simultaneamente ativos. Conforme a co-atividade continue

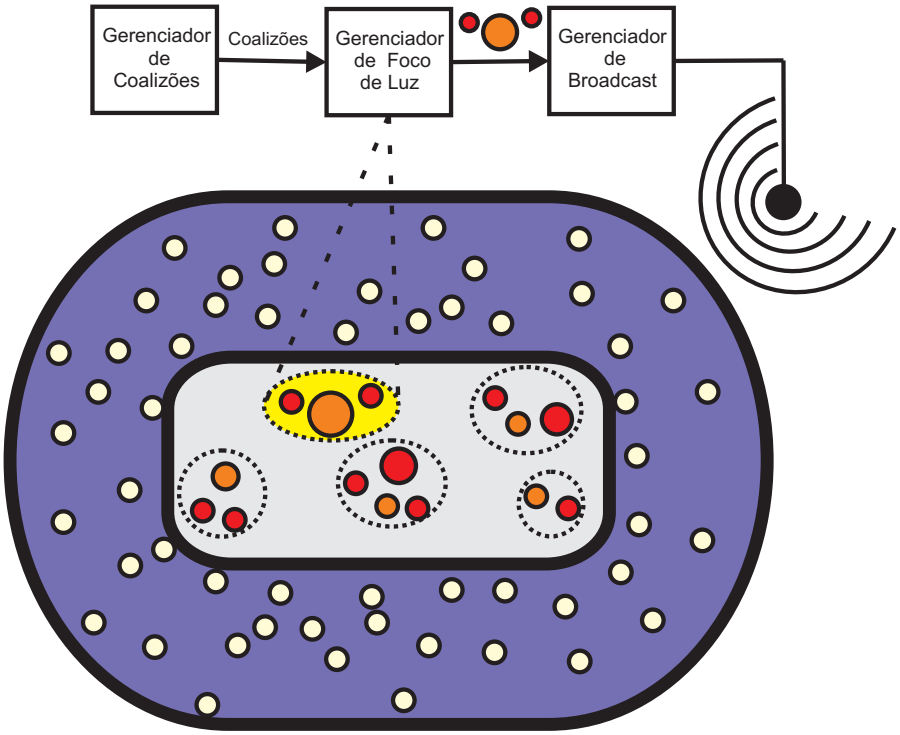


Figura 4.5: Mecanismo de Consciência. Em roxo temos a arquibancada e em cinza o campo. Os codelets inativos (círculos amarelos) estão na arquibancada, enquanto que os codelets ativos (círculos laranjas e vermelhos) estão no campo. O raio do círculo representa o seu grau de ativação. Círculos laranjas representam codelets de atenção. Na parte superior estão os componentes: gerenciador de coalizões, o qual calcula as coalizões presentes no campo; o gerenciador de foco de luz, que seleciona a coalizão de maior ativação; e o gerenciador de broadcast, que envia o conteúdo da coalizão escolhida a todos os codelets cadastrados.

existindo, essas *associações* se fortalecem, gerando coalizões de dois ou mais *codelets*. Um codelet pode estar presente em mais de uma coalizão.

4.8.4 Gerenciador de Foco de Luz

Várias coalizões podem ser formadas no campo da arena desportiva. Cada uma delas tem seu próprio nível de ativação, o qual é calculado pela média dos níveis de ativação dos codelets que a compõem. Esse nível de ativação determina o grau de competitividade de uma coalizão para entrar na "consciência". O *gerenciador de foco de luz* é o responsável por calcular o nível de ativação das coalizões e escolher a coalizão de maior ativação. A coalizão vencedora receberá o "foco de luz", como no teatro de Baars.

4.8.5 Gerenciador de Broadcast

Na teoria do *workspace* global, o conteúdo da "consciência" deve ser transmitido a todos os *codelets* do sistema. O *Gerenciador de Broadcast* é o responsável por isso, compilando todas as informações dos codelets da coalizão vencedora que foi escolhida pelo gerenciador de foco de luz. Essa informação é então disseminada para todos os codelets e cada codelet escolhe se a informação é relevante para si ou não.

4.9 Rede de Comportamentos

Os mecanismos de seleção de ação são fundamentais no desenvolvimento de arquiteturas de controle baseadas em comportamento. IDA utiliza a Rede de Comportamentos desenvolvida em (Maes, 1989), com aprimoramentos desenvolvidos em (Negatu & Franklin, 2002; Negatu, 2006). Esse mecanismo, juntamente com o módulo de consciência, formam o núcleo da arquitetura e são os componentes mais importantes.

A rede de comportamentos é uma rede distribuída, recorrente e não hierárquica. Ela agrupa um modelo de computação conexionista e uma estrutura de representação simbólica. Esse é o mesmo mecanismo encontrado nos agentes VMattie (Song, 1998; Song & Franklin, 2000), CMattie (Franklin & Graesser, 1999), IDA (Franklin, 2003; Negatu, 2006) e CTS (Dubois, 2007a)[11].

A rede de comportamentos de Maes foi construída diante do contexto apresentado na *Sociedade da Mente* (Minsky, 1986). Segundo Minsky, um sistema inteligente pode ser construído através de uma sociedade de agentes com competências específicas[12]. Assim, dado um conjunto de agentes bastante simples, é possível, através da interação entre eles, buscar um objetivo comum ou emergir um determinado comportamento.

A arquitetura Baars-Franklin seleciona e executa ações para atender aos seus propósitos internos. Podem existir diversos propósitos operando em paralelo cuja urgência varia com o tempo e com as mudanças do ambiente. Uma rede de comportamentos ativa é composta por estruturas chamadas de *cadeias de comportamentos* (veja figura 4.6). As cadeias de comportamento são compostas de *nós de comportamentos* e de *objetivos*. Em geral, os comportamentos são ações de nível médio, as quais dependem de vários *codelets de comportamento* para a sua execução. Os objetivos são similares aos comportamentos com a diferença de que suas ações são para concluir uma meta. Um comportamento corresponde a um *contexto de objetivo* da GWT (Negatu, 2006).

O mecanismo de seleção de ações de ABF funciona em conjunto com o mecanismo de "consciência". Uma cadeia de comportamentos é instanciada se tornando parte da rede de comportamentos ativa quando um *codelet de informação* encontra uma informação relevante no *broadcast* "consciente". Comportamentos, assim como os *contextos de objetivo* de Baars, cooperam e competem com outros na rede de comportamentos. A dinâmica da rede de comportamentos eventualmente seleciona um comportamento

[11] Mecanismos de seleção de comportamentos são amplamente estudados na literatura. Outros mecanismos de seleção de ação baseados em comportamentos podem ser encontrados em: (Brooks, 1986; Tyrrell, 1993; Blumberg, 1994; Low *et al.* , 2004; Pinto, 2005; González *et al.* , 2006; Ros *et al.* , 2009).

[12] Essa característica da "multiplicidade da mente" aproxima o modelo ao colocado em (Jackson, 1987) e se mostrou uma solução que se ajusta bem à teoria do *workspace* global de Baars.

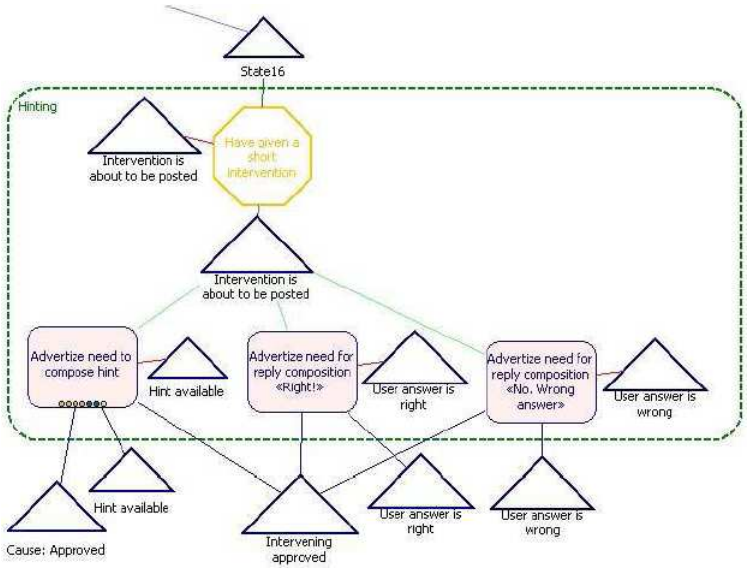


Figura 4.6: Exemplo de cadeia de comportamentos. Os retângulos representam os comportamentos, os triângulos proposições que fazem parte das pré-condições e pós-condições de um comportamento. Por fim, os hexágonos representam os objetivos. Fonte: (Dubois, 2007a).

relevante para entrar em ação, o qual entra no campo da arena desportiva e executa as suas atividades.

4.9.1 Estrutura da Rede

Os nós da rede representam competências pré-definidas simples como "buscar comida" ou "beber água", os quais possuem uma medida de ativação. A característica de seleção dinâmica e distribuída de comportamentos vem de duas ondas de propagação de ativação. Uma fonte de ativação são os estímulos externos providos pelos sensores do ambiente. Esses sensores verificam se uma determinada proposição é verdadeira ou falsa. Uma outra fonte são as motivações que são geralmente derivadas de estímulos internos. As motivações possuem valores reais entre "0" e "1". Essas ondas de ativação criam a capacidade de planejamento (implícito) ativando mais a sequência de comportamentos, que leva o estado atual ao estado desejado através da relevância da situação corrente (ativação forward) e em relação aos objetivos traçados (ativação backward).

Cada nó (ver figura 4.8) representa um comportamento[13]. Um nó i é descrito pela ênupla $(c_i, a_i, d_i, \alpha_i)$, em que: c_i é uma lista de pré-condições (proposições) que devem ser satisfeitas antes do nó se tornar ativo. a_i e d_i representam os efeitos esperados da ação realizada pelo comportamento, sendo o primeiro a lista de proposições que possivelmente se tornarão verdadeiras com a execução do comportamento (lista de adição) e o segundo a lista de proposições que se tornarão falsas (lista de remoção). α_i é o nível de ativação do comportamento. Um nó também possui um código executável que

[13] Maes (1989) chama de módulo de competência.

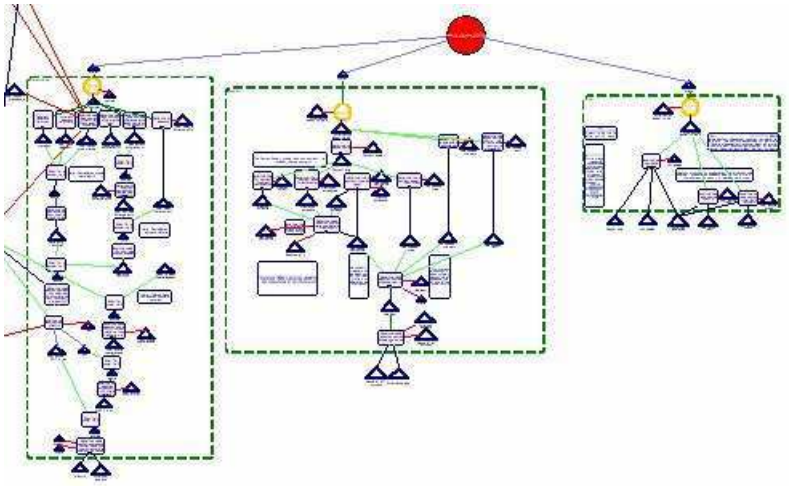


Figura 4.7: Parte da rede de comportamentos de *CTS*. Várias cadeias (dentro dos retângulos pontilhados) formam a rede de comportamentos. Fonte: (Dubois, 2007a).

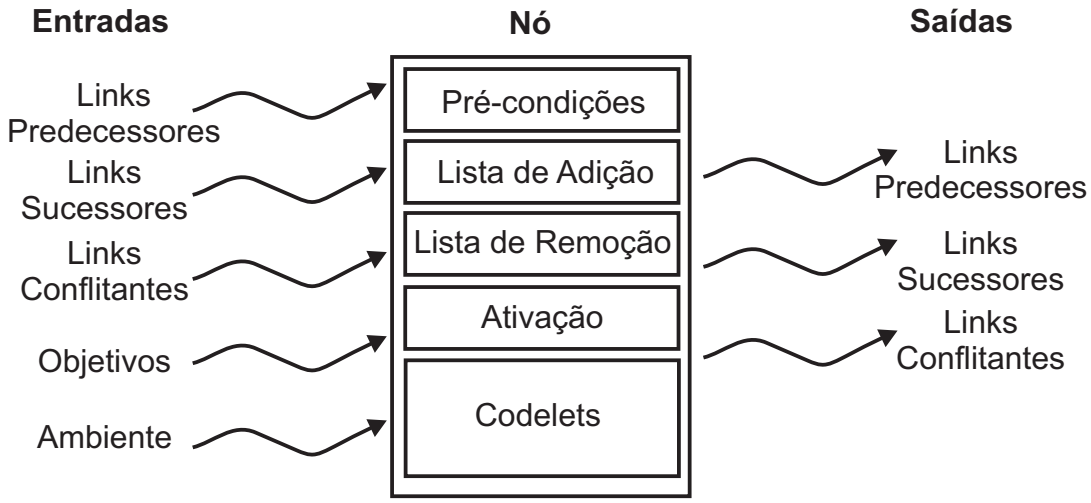


Figura 4.8: Nó da Rede de Comportamentos.

efetivamente realiza as ações ligadas ao comportamento. Esse código, implementado em *IDA* através de *codelets*, pode estar ligado diretamente com o controle, fazer inferências ou qualquer outra operação relevante ao comportamento.

Os comportamentos são interligados por *links* (internos) causais, que servem para a propagação da ativação entre os nós. Há três tipos de *links* internos: *sucessores*, *predecessores* e *conflitantes*. Eles são criados através da verificação do compartilhamento entre as proposições dos comportamentos nas listas de pré-condições, de adição e de remoção. Formalmente, os *links* são descritos a seguir:

- Há um *link sucessor* de um nó x para um nó y ("x tem y como sucessor") para cada proposição p, que pertence à lista de adição de x e também à lista de pré-condições de y (isso permite que exista mais de um *link* sucessor entre dois nós). De maneira formal: dado um nó $x = (c_x, a_x, d_x, \alpha_x)$ e um nó $y = (c_y, a_y, d_y, \alpha_y)$,

há um *link* sucessor de x para y, para cada proposiçao $p \in a_x \cap c_y$.

- Um *link predecessor* de um nó x para um nó y ("x tem y como predecessor") existe para cada *link* sucessor de y para x. Formalmente, dado um nó $x = (c_x, a_x, d_x, \alpha_x)$ e um nó $y = (c_y, a_y, d_y, \alpha_y)$, há um link *predecessor* de x para y, para cada proposição $p \in c_x \cap a_y$.

- Existe um *link conflitante* de um nó x para um nó y ("y conflita com x") para cada proposição p, que é membro da lista de remoção de y e faz parte da lista de pré-condições de x. Formalmente, dado um nó $x = (c_x, a_x, d_x, \alpha_x)$ e um nó $y = (c_y, a_y, d_y, \alpha_y)$, há um *link* conflitante de x para y, para cada proposição $p \in c_x \cap d_y$.

Além desses *links* internos, há *links* externos que são responsáveis por propagar ativações vindas do ambiente. São eles:

- *links do ambiente* - dado um nó $x = (c_x, a_x, d_x, \alpha_x)$, se a proposição p sobre o ambiente é verdadeira e a proposição $p \in c_x$ (isto é, x é parcialmente apropriado ao estado atual), então há um link ativo (excitatório) do sensor da proposição p para o nó x.

- *links dos objetivos* - dado um nó $x = (c_x, a_x, d_x, \alpha_x)$, se um objetivo g tem ativação maior que zero e o objetivo $g \in a_x$ (ou seja, se x pode satisfazer o objetivo g), então há um link ativo (excitatório) do objetivo g para o nó x.

- *links dos objetivos protegidos* - dado um nó $x = (c_x, a_x, d_x, \alpha_x)$, se um objetivo g tem um nível de ativação maior que zero e o objetivo $g \in d_x$ (isto é, se x desfaz g, ou impossibilita que g seja atingido), então há um link ativo (inibidor) do objetivo g para o nó x.

4.9.2 O mecanismo

Um comportamento é *executável* em um tempo t quando todas as suas pré-condições são verdadeiras no tempo t. Um comportamento *executável* pode ser selecionado para executar ações pela execução de seu código executável, quando seu nível de ativação ultrapassar um determinado limiar global. Se mais de um comportamento é executável após um ciclo, então o que possuir maior ativação é escolhido. O algoritmo global executa a cada passo, a fim de selecionar um novo comportamento, o seguinte laço:

1. Ambiente e objetivos têm a sua excitação calculada e é injetada a ativação nos comportamentos. A inibição vinda dos objetivos protegidos também é computada e retira ativação dos comportamentos.

2. Ativação e inibição através dos links sucessores, predecessores e conflitantes são calculadas

3. Os níveis de ativação são normalizados. Assim o nível de ativação médio permanece constante.

4. Um comportamento é ativo se: *(i)* é executável; *(ii)* seu nível de ativação é maior que um limiar; *(iii)* caso tenha o maior nível de ativação dentre os outros que satisfazem *(i)* e *(ii)*.

5. O comportamento selecionado, após realizar suas ações, tem seu nível de ativação ajustado para zero e o limiar global é colocado no valor padrão.

6. Se nenhum comportamento foi selecionado, reduz-se o limiar em 10%.

O algoritmo de Maes possui parâmetros globais que afetam diretamente a propagação da ativação e podem ser utilizados para dar determinadas características à execução. São eles: π é o valor médio da ativação após cada passo (usado na normalização); θ é o limiar para um comportamento se tornar ativo; ϕ é a quantidade de energia injetada por cada proposição que é dada como verdadeira no ambiente; γ é a quantidade de ativação que é injetada por um objetivo global; δ é a quantidade de ativação que um objetivo protegido retira da rede.

O cômputo da ativação propagada internamente é calculado da seguinte maneira:

- *excitação dos sucessores* - um nó executável $x = (c_x, a_x, d_x, \alpha_x)$ envia ativação adiante (uma fração do seu próprio nível de ativação) para um nó $y = (c_y, a_y, d_y, \alpha_y)$, através de um *link* sucessor, se esse *link* corresponde a uma proposição $p \in a_x \cap c_y$ que é falsa. A quantidade de ativação é dada por $\frac{\alpha_x \phi}{\gamma N M}$, em que N é o número de proposições em c_y e M é o número de nós que possuem a proposição p em suas listas de pré-condições.

- *excitação dos predecessores* - um nó $x = (c_x, a_x, d_x, \alpha_x)$, que não está executável, enviará uma ativação para um nó $y = (c_y, a_y, d_y, \alpha_y)$, através de um *link* predecessor, se o *link* corresponde a uma proposição $p \in c_x \cap a_y$ que é falsa. A quantidade de ativação é dada por $\frac{\alpha_x}{NM}$, em que N é o número de proposições em a_y, e M é o número de nós que tem p em suas listas de adição.

- *inibição dos conflitantes* - Cada nó $x = (c_x, a_x, d_x, \alpha_x)$, executável ou não, enviará uma parcela de inibição através dos *links* conflitantes para um nó $y = (c_y, a_y, d_y, \alpha_y)$, se o *link* corresponder a uma proposição $p \in c_x \cap d_y$ que é verdadeira. A quantidade de ativação é dada por $\frac{\alpha_x \delta}{\gamma N M}$, em que N é o número de proposições em d_y e M é o número de nós que têm a proposição p em suas listas de remoção. O nó x irá remover ativação de y somente se $\alpha_x > \alpha_y$.

O cálculo de ativação vinda do ambiente e dos objetivos é calculado como a seguir:

- excitação do ambiente - se há um *link* entre um nó $x = (c_x, a_x, d_x, \alpha_x)$ e o ambiente através de uma proposição p, então a quantidade $\frac{\phi}{NM}$ de ativação é injetada no nó x, onde N é o número de proposições em c_x e M é o número de nós que têm a proposição p em suas listas de pré-condição.

- excitação vinda dos objetivos - se há um *link* entre um nó $x = (c_x, a_x, d_x, \alpha_x)$ e um objetivo devido a uma proposição p, então a quantidade $\frac{\gamma}{NM}$ é injetada, onde N é o número de proposições em a_x e M é o número de nós que podem atingir p (possuem p em suas listas de adição).

- inibição vinda dos objetivos protegidos - se um nó $x = (c_x, a_x, d_x, \alpha_x)$ pode desfazer um objetivo g então a quantidade $\frac{\delta}{NM}$ é removida dele, onde N é o número de proposições em d_x e M é o número de comportamentos que podem desfazer o objetivo g.

4.10 Ciclo Cognitivo

A arquitetura Baars-Franklin possui um ciclo de operação, chamado de *ciclo cognitivo*, que foi apresentado inicialmente em (Baars & Franklin, 2003). Esse ciclo resume algumas hipóteses de Baars e Franklin sobre a cognição humana. Ele é responsável pelo processamento da arquitetura e envolve a contínua interação entre os módulos de percepção, memória de trabalho, memória episódica, memória associativa de longo prazo, consciência, mecanismo de seleção de ação, e, atuação propriamente dita[14].

A ideia de "ciclo" ou, mais explicitamente, de uma sequência de fenômenos ocorrendo em uma ordem determinada, pode parecer antagônica com os vários módulos interagindo paralelamente e de forma independente como traz a teoria do *workspace* global. No entanto, um ciclo se faz necessário uma vez que a arquitetura *multi-thread* - o que estaria mais próximo de um paralelismo em uma arquitetura serial - acaba sendo por demais custosa, e leva a problemas de sincronismo de difícil solução. O "ciclo" cognitivo , portanto, tem a finalidade de facilitar a implementação computacional, sem prejudicar as principais ideias da proposta original de Baars.

O ciclo cognitivo de IDA[15] possui nove estágios distribuídos em três partes: percepção, interpretação e ação. Os estágios 1-3 envolvem percepção, memória associativa de longo prazo, memória episódica e a memória de trabalho da ABF (Ramamurthy *et al.* , 2004). Os estágios de 4-9 correspondem à interação da "consciência" (Bogner, 1999) e os módulos de seleção de ação (Negatu & Franklin, 2002; Negatu, 2006). Considerando que os processos de cognição e consciência humanas se dão em um fluxo contínuo, como uma sucessão de episódios, segundo (Crick & Koch, 2003), os humanos estão constantemente tendo sensações, sem esperar que um ciclo de percepção tenha sido completamente processado. Nesse sentido, Baars & Franklin (2003) defendem que é possível uma sobreposição entre os passos do ciclo cognitivo, o que permitiria tomar decisões e ações em paralelo: "Nós conjecturamos que um ciclo cognitivo completo deva levar no mínimo 200 ms. Mas devido à sobreposição e automaticidade, as quais reduzem o ciclo, é possível ter vinte ciclos rodando por segundo."

Essa diminuição do tempo de um ciclo acontece pela criação de automatismos (conjuntos de ações pré-estabelecidas para a obtenção de um resultado), que reduzem os

[14] (Dubois, 2007a, p. 85) defende que o ciclo cognitivo pode ser comparado com o mecanismo de reentrância de Edelman (veja 2.3.3). Segundo o princípio de reentrância, em um processo dinâmico e contínuo, grupos neurais e mapas entre esse grupos trocam estímulos (excitatórios e inibitórios) até que um padrão estável suficientemente forte surja e se torne consciente. De maneira similar, isso ocorre na arquitetura Baars-Franklin: o broadcast incluso no ciclo propaga informações aos recursos inconscientes para que esses respondam aos estímulos vindos de outros ou do ambiente externo. Contudo, para acontecer um *broadcast*, um padrão já deve ter se estabilizado na memória de trabalho anteriomente.

[15] Dubois (2007a), com o seu agente CTS, traz uma versão desse ciclo cognitivo ligeiramente modificada do ciclo de *IDA*.

broadcasts (passo 5), que são seriais e possuem uma capacidade limitada de processamento, de acordo com a teoria do *workspace* global. Negatu (2006) defende que os automatismos aumentam a sobreposição dos ciclos cognitivos, incrementando o paralelismo e evitando o uso da "consciência" no desempenho de atividades automatizadas[16].

Os nove passos do ciclo cognitivo de IDA são descritos a seguir (adaptado de (Baars & Franklin, 2003)):

(1) Percepção

Estímulos sensoriais (externos e internos) são recebidos e interpretados pela percepção. Esse estágio pode ser dividido em duas fases: a primeira, chamada de *percepção precoce*, em que novas entradas são percebidas pelos sensores e codelets especializados, que encontram características relevantes para sua especialidade, ativam os nós apropriados da Slipnet. A segunda fase, chamada de *percepção de chunks*, acontece através da ativação que passa de nó para nó na Slipnet. A Slipnet se estabiliza e converge sequências de chunks menores em chunks maiores. Esses chunks maiores, representados por nós de significado na Slipnet, constituem um *percepto*.

(2) Percepto no *buffer* pré-consciente

O percepto, incluindo dados com significado, é armazenado no *buffer pré-consciente* da memória de trabalho de IDA. Esses *buffers* podem envolver informações visuo-espaciais, fonológicas, entre outras.

(3) Associações locais

A memória episódica transiente e a memória associativa de longo prazo são acessadas, utilizando os perceptos e o conteúdo residual do *buffer* pré-consciente como sinal de entrada. O conteúdo do *buffer* pré-consciente adicionado ao conteúdo recuperado das associações locais forma a memória episódica transiente e a memória associativa de longo prazo.

(4) Competição pela consciência

Codelets de atenção, cuja função é trazer eventos relevantes, urgentes ou insistentes para a consciência, observam a memória de trabalho de longo prazo. Alguns extraem informações, formam coalizões e ativamente competem por acesso à consciência. A competição pode incluir também codelets de atenção de ciclos passados recentes. A ativação de um codelet de atenção decai rapidamente, dificultando a competição com codelets de atenção que estão competindo pela primeira vez. Entretanto, o conteúdo de coalizões sem sucesso é mantido no *buffer* pré-consciente e pode auxiliar em decisões, no caso de perceptos ambíguos.

(5) Broadcast consciente

Uma coalizão de codelets, tipicamente um codelet de atenção e os codelets de informação relacionados, ganha acesso ao workspace global e tem o seu conteúdo propagado

[16] ver (Negatu, 2006, Cap. 5) para questões sobre automatização e de-automatização.

para todos os outros codelets. Hipoteticamente, esse broadcast corresponde à consciência fenomenal. O conteúdo atual da consciência também é armazenado na memória episódica transiente. De tempos em tempos (não relacionado com o ciclo cognitivo), o conteúdo da memória episódica transiente é consolidado na memória associativa de longo prazo. A memória episódica transiente é também uma memória associativa com uma taxa de decaimento na ordem de horas. Isso a distingue de uma memória autobiográfica, uma parte da memória associativa de longo prazo.

(6) Recrutamento de recursos

Codelets de comportamento relevantes, ou seja, codelets cujas variáveis podem ser associadas às informações do broadcast, respondem ao broadcast consciente. Se o codelet de atenção vencedor for um codelet de expetativa apontando para um resultado inesperado de uma ação anterior, os codelets que respondem são aqueles que auxiliam a sair da situação inesperada. Assim, a consciência resolve o problema de relevância no recrutamento de recursos.

(7) Configurando a hierarquia de contextos de objetivos

Alguns codelets de comportamento ativos instanciam uma cadeia de comportamentos apropriada, caso a cadeia atual não seja adequada. Eles também associam variáveis e enviam ativação aos outros comportamentos. Nesse estágio há a premissa de que existe pelo menos um codelet de comportamento executável e uma cadeia de comportamentos instanciada. Caso contrário, uma rotina de solução de problemas é evocada utilizando mecanismos adicionais.

(8)Escolha da ação

A rede de comportamentos escolhe um único comportamento e o executa. Essa escolha pode vir de uma cadeia de comportamentos que acabou de ser instanciada ou de uma previamente ativa. Essa escolha é afetada pelas motivações internas (ativação vinda dos objetivos) pela situação atual, pelas condições internas e externas e pela relação entre os comportamentos e seus valores de ativação.

(9) Realização da ação

A execução do comportamento resulta na realização das atividades especializadas pelos codelets de comportamento associados ao comportamento escolhido. Isso pode ter consequências internas e externas e é quando o *IDA* está entrando em ação. Os codelets de comportamento ativos possuem um codelet de expectativa que monitora as ações realizadas e tenta trazer à consciência qualquer desvio dos resultados esperados.

4.11 Interação da Rede de Comportamentos e do Mecanismo de Consciência

Apesar de estarem implementados em módulos a parte, a rede de comportamentos e o mecanismo de consciência estão ligados e dependem um do outro para o correto funcionamento do agente.

No sentido da rede de comportamentos para o mecanismo de consciência, um dos pontos de interação são os próprios codelets de comportamento que nas implementações de *IDA* e *CTS* são sempre adicionados ao campo de jogo[17]. Isso permite que esses codelets possam participar de coalizões e possivelmente alterar o resultado do *broadcast* consciente. Outro ponto é que os codelets de comportamento podem também gerar codelets de informação que em rodadas futuras do ciclo cognitivo se tornem relevantes e modifiquem o broadcast consciente.

No sentido oposto, os *broadcasts* da consciência podem de alguma forma alterar o estado das proposições da rede de comportamentos ou codelets podem alterar os objetivos, reduzindo ou aumentando o seu peso na rede de comportamentos ou mesmo remover ou adicionar novos objetivos temporariamente.

4.12 Conclusão

Esse capítulo finaliza os estudos teóricos de consciência artificial. A arquitetura Baars-Franklin difere das outras abordagens apresentadas no capítulo 3, pois se baseia na teoria do *workspace* global de Baars e apresenta um conceito procedimental de consciência. Além disso, um dos propósitos principais da arquitetura é a implementação computacional do modelo de Baars. Essa abordagem, traz um maior embasamento no desenvolvimento da arquitetura, removendo grande parte da ingenuidade contida nas outras abordagens.

Um ponto muito interessante da arquitetura Baars-Franklin é a integração de um sistema paralelo (inconsciente) formado pelos codelets e um sistema serial gerado pelo mecanismo de foco de luz da consciência. Essa integração se assemelha a proposta de Dennett de que a consciência pode ser entendida como *"a operação de uma máquina virtual 'a la von Neumann' implementada em uma arquitetura paralela do cérebro."*

No próximo capítulo é mostrado uma implementação da arquitetura Baars-Franklin para um problema de uma criatura virtual em um ambiente simulado, permitindo a realização de experimentos e análises de vantagens e desvantagens do uso dessa abordagem de consciência artificial.

[17] Da metáfora da arena de (Jackson, 1987).

Capítulo 5

Problema Exemplo: Navegação de Veículo Autônomo

> There are those who look at things the way they are, and ask why. I dream of things that never were, and ask why not?
>
> Robert Kennedy

5.1 Introdução

Esse capítulo apresenta um exemplo de aplicação da arquitetura Baars-Franklin com o intuito de verificar as implicações do uso dessa tecnologia em criaturas artificiais. Para isso, foi escolhido um problema de navegação autônoma - problema exemplo distinto das outras aplicações onde a ABF foi empregada - para a realização das simulações e para ilustrar as análises da arquitetura computacional.

O problema que foi abordado nesse trabalho é o de um veículo autônomo (criatura artificial) que se encontra em um ambiente virtual composto por obstáculos pré-definidos pelo usuário. Problemas dessa natureza tem sido constantemente tratados pela literatura por meio das mais diversas abordagens e soluções de controle (Gudwin, 1996; de Toro, 2007; Fernandez-Leon *et al.* , 2009; Hui & Pratihar, 2009; McFetridge & Ibrahim, 2009).

Na seção 5.2, é apresentada a abordagem utilizada no problema de navegação autônoma. A seção 5.3, explica a arquitetura do *CAV*, o agente autônomo "consciente" implementado para tratar o problema de navegação autônoma. Nessa seção também são feitas considerações qualitativas e quantitativas sobre a arquitetura implementada.

5.2 O problema exemplo

O problema exemplo de navegação autônoma utilizado nesse trabalho foi inicialmente modelado em (Gudwin, 1996) e revisitado em (Suárez, 2000; de Toro *et al.* ,

2007). Partindo-se das mesmas abstrações adotadas anteriomente, mas com algumas melhorias no ambiente de simulação, buscou-se ilustrar as vantagens e desvantagens da aplicação de técnicas de consciência artificial no controle de um veículo autônomo.

No ambiente de simulação proposto por (Gudwin, 1996), o veículo autônomo se movimenta por um mundo virtual repleto de objetos variados. Alguns objetos podem alterar a energia armazenada nas baterias do veículo através do contato físico. O veículo deve conseguir manter a carga de sua bateria em um nível operacional aceitável. Além das questões energéticas, o veículo recebe uma posição no ambiente para a qual ele deve se dirigir, chamada de *meta*, e deve ser capaz de se mover no ambiente entre os obstáculos para atingir a meta estipulada, sem colidir com os obstáculos.

5.2.1 Descrição do veículo: Sensores e Atuadores

O veículo possui sensores que lhe permite observar o seu estado interno (*sensor de carga de bateria*) e sensores que possibilitam extrair informações do ambiente (*sensores de informação remota* e *sensores de contato*).

O controle do veículo pode ser feito através de seus atuadores, que são capazes de determinar: a posição do sensor de informação remota, o eixo longitudinal e a velocidade nominal do veículo. Os sensores e atuadores são detalhados a seguir.

Sensores de Informação Remota

Os sensores de informação remota são uma abstração de um mecanismo de visão robótica. São formados por uma matriz de (7x7)[1] sensores capazes de cobrir uma área quadrada e detectar cores e informação de posição dos objetos à distância.

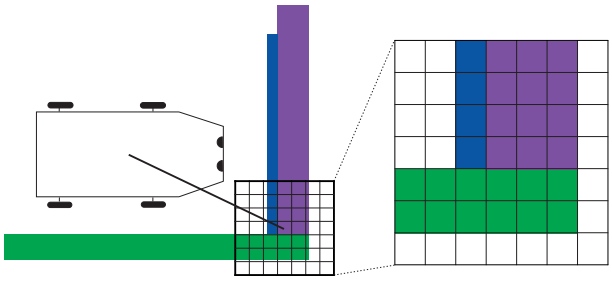


Figura 5.1: Sensores de Informação Remota. Os sensores totalmente ou parcialmente cheios são ativados pelo obstáculo observado.

[1] Gudwin (1996) originalmente propõe um modelo com uma matriz (8x8). Como foi necessária uma nova implementação das funções de tratamento desses sensores, optou-se arbitrariamente por um novo modelo 7x7. Essa alteração não traz mudanças significativas ao modelo original.

Sensores de Contato

O veículo possui quatro sensores de contato, localizados nas extremidades do veículo, que são fixos. Ao serem estimulados por um objeto, esses sensores podem perceber as características do objeto observado.

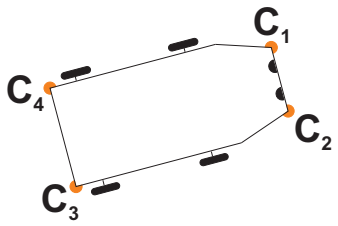


Figura 5.2: Sensores de contato. c_1, c_2, c_3, c_4 são sensores afixados às extremidades do veículo.

Sensor de Carga de Baterias

Esse sensor possibilita verificar a carga das baterias em termos percentuais. Se as baterias estiverem completamente descarregadas, o sensor acusará "0" e no extremo oposto, caso estejam totalmente cheias o sensor marcará "100". A descarga da bateria é diretamente proporcional ao módulo da velocidade do veículo. Caso o veículo esteja parado, mas ligado, a descarga é lenta e linear no tempo. Quando o veículo entra em contato com um objeto que pode lhe dar energia há uma carga rápida.

Atuadores de posição dos sensores de informação remota

A posição dos sensores de informação remota é determinada pelo ângulo ϕ, medido a partir da direção do veículo e o raio de ação ρ, calculado a partir do centro do carro. Através desses atuadores, os sensores de informação remota podem ser colocados em qualquer posição do ambiente, respeitando uma distância máxima do veículo e um ângulo máximo.[2].

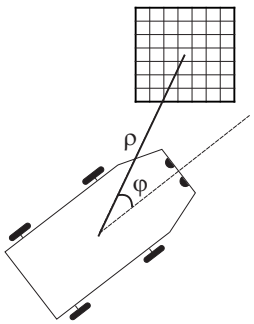


Figura 5.3: Atuadores do sensor de informação remota. Os sensores de informação remota podem ser posicionados através dos atuadores ρ e ϕ.

[2]Na implementação desse trabalho $0 \leq \rho \leq 150$ em relação ao centro do veículo e $-60° \leq \phi \leq 60°$ em relação à direção do veículo.

Atuadores de Movimentação do Veículo

O modelo proposto oferece dois atuadores para movimentação do veículo. Um é o atuador sobre a velocidade nominal v do veículo. Valores negativos de v correspondem a dar ré, v positivo significa ir para frente e para v igual a zero o veículo está parado. O outro atuador de movimentação diz respeito ao ângulo θ da posição das rodas em relação ao eixo longitudinal do veículo. Ele atua na direção que o veículo segue, caso $v \neq 0$[3].

5.2.2 Ambiente de Navegação

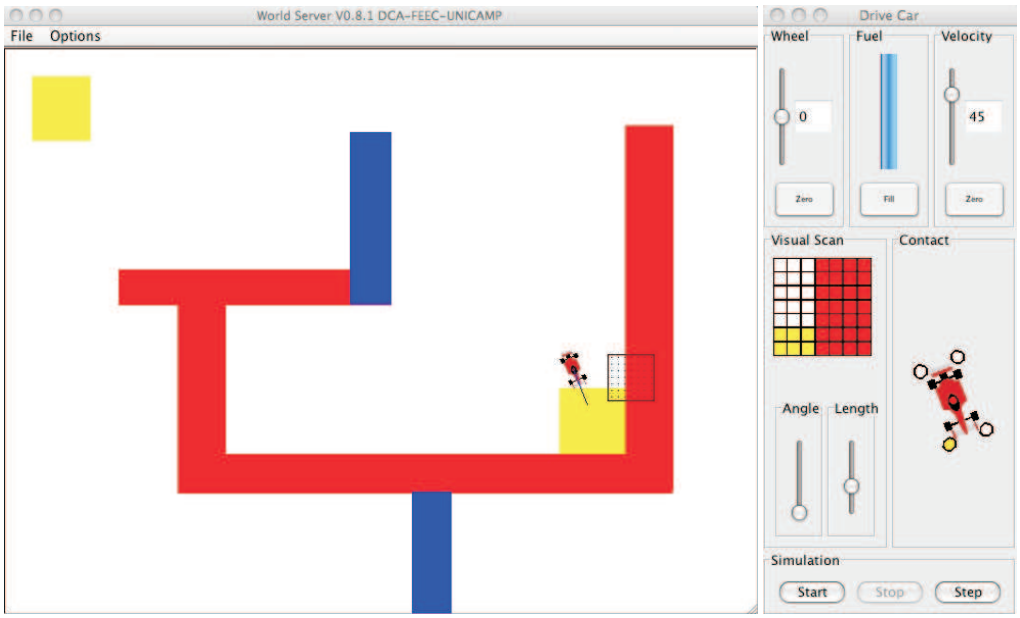


Figura 5.4: Ambiente de simulação. À esquerda, a janela principal em que é criado o mapa do mundo virtual. Através dela também é possível observar o andamento do veículo durante a simulação. À direita, a janela de acompanhamento dos sensores e atuadores.

O ambiente em que o veículo navega é composto por um retângulo cercado de paredes e que contém objetos variados em seu interior. Os objetos podem ser caracterizados por três propriedades físicas: cor, dureza e gosto. A cor serve para identificar visualmente os objetos. Por meio da dureza é possível determinar se o veículo pode ou não navegar sobre esses objetos. Quando dureza é igual a "1", os objetos são denominados

[3]Nesse trabalho $-45° \leq \theta \leq 45°$. Quando $\theta = 0$ o carro não muda a sua direção. Valores negativos correspondem a virar a roda para a esquerda e positivos para a direita.

intransponíveis, no caso de ser "0", é dito que os objetos são *transponíveis*[4]. O gosto é detetado pelo sensor de contato e varia de "-1" a "1", sendo "-1" equivalente a desprazer, "1" equivalente a prazer e "0" à indiferença[5]. A transferência de energia corresponde à propriedade de certos objetos de fornecer ou absorver energia do veículo durante o contato.

O ambiente de simulação utilizado (ver figura 5.4) vem sendo desenvolvido pelo Grupo de Pesquisa em Cognição Artificial (GRACO) do DCA/FEEC/UNICAMP. Esse simulador permite ao usuário construir diversos ambientes de teste, através da alteração das características e do posicionamento dos objetos. Historicamente, ele tem evoluído com passar dos anos: o primeiro ambiente foi desenvolvido na linguagem C por (Gudwin, 1996), implementado em uma arquitetura cliente-servidor por (Suárez, 2000) e reimplementado em Java em (de Toro, 2007). Atualmente, encontra-se em desenvolvimento uma versão 3D do simulador.

Nesse trabalho, foram adicionadas mudanças conceituais importantes ao simulador. Diferente de (de Toro, 2007), na arquitetura cliente-servidor, o simulador faz o papel de servidor, enquanto que o cliente é o controlador (ver figura 5.5). Uma outra mudança fundamental é a característica assíncrona entre a comunicação do controlador e do simulador; não há um ponto no ciclo de simulação em que é esperado um comando do controlador. Os comandos do controlador podem ser enviados a qualquer momento. O simulador continua com a simulação em andamento, independente da recepção de novos comandos. Nos trabalhos anteriores (Gudwin, 1996; de Toro, 2007) ocorria uma sincronização de um passo de simulação para cada passo do controlador. Essa mudança foi realizada por tornar a simulação mais parecida com o que ocorre no mundo real.

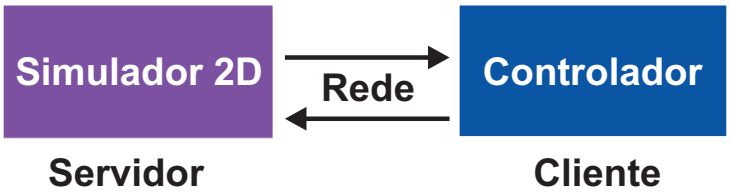


Figura 5.5: Arquitetura cliente-servidor do ambiente de simulação. O servidor é o ambiente de simulação e o cliente é um controlador para o veículo autônomo.

[4] Gudwin (1996) propõe que a velocidade é alterada pela característica dureza que um objeto pode ter. Nos experimentos desse trabalho, ou é possível passar por um objeto (no caso de ser um ponto de recarga), ou o objeto é completamente intransponível, como uma abstração de uma parede por exemplo.

[5] Essa característica foi desprezada nos experimentos desse trabalho.

5.3 CAV: O agente autônomo consciente

Para atender o problema exemplo de navegação autônoma foi desenvolvido um agente autônomo consciente[6], denominado *CAV (Conscious Autonomous Vehicle)*. *CAV* faz uso de parte da tecnologia *IDA* desenvolvida por Stan Franklin. Foi realizado um convênio entre a UNICAMP e a Universidade de Memphis que disponibilizou a implementação do framework ConAg (Bogner, 1999) e de uma implementação da rede de comportamentos com as modificações sugeridas em (Negatu, 2006).

Durante a implementação de CAV algumas classes tiveram suas estruturas de dados atualizadas, foi criada uma rede de comportamentos em XML[7] e diversas adaptações foram realizadas para o funcionamento em conjunto da consciência com a rede de comportamentos. Essas adaptações serão explicitadas no decorrer da explicação da arquitetura proposta.

5.3.1 Arquitetura de CAV

CAV é uma arquitetura cognitiva que possui módulos de memória de trabalho, consciência, rede de comportamentos, implementados de maneira distribuída pelos codelets (ver figura 5.6). O escopo de implementação desse agente não contemplou os módulos de memória de longo prazo e episódica, metacognição e emoções. O enfoque dado nesse estudo está na criação da arquitetura cognitiva e que seus módulos e codelets são passíveis de aprimoramentos e pesquisas futuras para a implementação satisfatória em condições reais (como o módulo de percepção por exemplo).

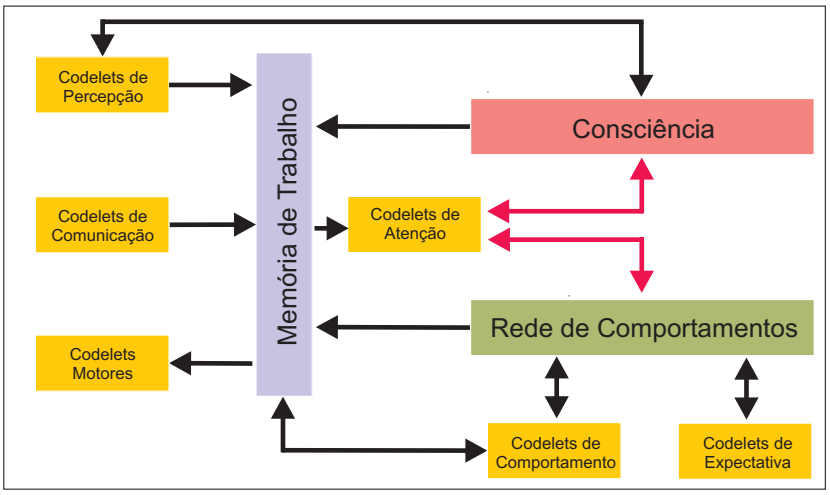


Figura 5.6: Arquitetura *CAV*

[6] Franklin (2003) faz diversas ponderações sobre o questionamento se *IDA* seria fenomenologicamente consciente e conclui que não há argumentos plausíveis para se considerar isso. Nesse trabalho, apesar de *consciente* não estar grafado entre aspas como sugere Stan Franklin, não há a motivação de implementar a consciência fenomenal como vista nos seres humanos. O agente é dito *consciente* por se inspirar em teorias de consciência.

[7] Conforme especificado em (Negatu, 2006).

5.3.2 Codelets

Os codelets, assim como em IDA, são compreendidos como agentes no sentido de (Franklin, 1997), pois, apesar de simples e bastante especializados, possuem características de autonomia, percepção, processo e ação, contando também com uma agenda interna própria.

CAV, multi-agentes, comporta vários tipos de codelets, os quais podem ser agrupados dependendo da sua função. Existem os codelets que ficam executando durante todo o funcionamento do agente e também outros que se mantém por um curto prazo.

Um ponto positivo da arquitetura Baars-Franklin é que sua estrutura oferece ao projetista uma grande liberdade para adicionar codelets conforme a necessidade da solução que se deseja propor. Por outro lado, toda essa liberdade, aliada à falta da formalização teórica da arquitetura pode significar uma dificuldade inicial no momento de se projetar o sistema. Para facilitar, há uma taxonomia de codelets como apresentada anteriormente[8]. Esse conjunto tem sofrido constantes alterações durante as pesquisas do grupo de Franklin. Além disso, não é incomum surgirem propostas de novos tipos de codelets, nem sempre ligados com uma teoria sobre a mente e seu funcionamento, mas relacionados somente ao domínio do problema.

Na proposta atual de CAV, os *codelets de informação* presentes em IDA (Franklin, 2005) e CTS (Dubois, 2007a) não foram utilizados pois no problema de navegação proposto os sinais vindos dos sensores do veículo foram armazenados diretamente na memória de trabalho e não necessitam de interpretação, como no caso das entradas textuais dos agentes anteriores. Por outro lado, outros tipos de codelets foram propostos. Os codelets de CAV podem ser divididos em codelets de comunicação, codelets de percepção, codelets de atenção, codelets de expectativa, codelets de comportamento e codelets motores. O funcionamento de cada codelet é explicado abaixo e um resumo pode ser encontrado na tabela 5.1.

Tabela 5.1: Tipos de codelets de *CAV*

Tipo	**Papel**
Comunicação	Realiza a comunicação com o simulador
Percepção	Interpreta o que o agente recebe do ambiente e armazena na memória de trabalho
Atenção	Monitora a memória de trabalho e rede de comportamentos, em busca de padrões esperados
Expectativa	Verifica se determinado resultado esperado na execução de um comportamento aconteceu
Comportamento	Altera os parâmetros dos codelets motores
Motor	Gera o comportamento de fato pela ação nos atuadores do veículo

[8] ver seção 4.4.

Codelets de Atenção

Os codelets de atenção são responsáveis por monitorar o estado interno do veículo ou detectar algum padrão interessante vinda dos sensores. Alguns codelets buscam por proposições verdadeiras no estado da rede de comportamentos, outros procuram por determinadas situações como a redução do nível das baterias ou a colisão com os obstáculos. Em *CAV* os codelets de atenção tem longa duração, ou seja, eles permanecem em funcionamento durante toda a execução do programa.

Codelets de Expectativa

Os codelets de expectativa são um tipo especial de codelets de atenção. Esses codelets têm vida curta e são gerados para acompanhar um comportamento específico. Eles monitoram o sistema para averiguar se o comportamento teve o resultado esperado durante a atuação. Se tudo aconteceu como programado, nada é preciso fazer, então ninguém é avisado disso. Por outro lado, passado um determinado tempo, caso o resultado não tenha sido encontrado, o codelet de atenção sinaliza que algo não aconteceu como esperado.

Codelets de Percepção

Os codelets de percepção são responsáveis por interpretar alguns sinais vindos do ambiente e armazená-los na memória de trabalho. Assim como em *IDA* e *CTS* os codelets de percepção em *CAV* se perpetuam durante a execução do agente. Os codelets de percepção também estão ligados a transporte de algumas variáveis, como em *IDA* e *CTS*. Nessa categoria há dois codelets com as características de interpretação de informações: os codelets que mantêm os obstáculos na memória de trabalho e os codelets que mantêm os *landmarks* do ambiente. Esses codelets, ao receberem um dado novo, realizam um processamento para armazenar na memória de trabalho na forma de informação útil para o agente. Por exemplo, o codelet que mantem os obstáculos na memória processa os sinais dos sensores remotos com a finalidade de adicionar alterações no modelo de mundo atual.

Codelets de Comunicação

Os codelets de comunicação cuidam da comunicação entre o servidor e o cliente, realizando a comunicação através de *sockets*. Esse codelet está ativo constantemente durante a execução do agente e atualiza de tempos em tempos o status do veículo.

Codelets de Comportamento

Os codelets de comportamento são os codelets que alteram os parâmetros da memória de trabalho utilizados pelos codelets motores. Esses codelets têm o conhecimento de *"o que fazer"* e representam o "código executável" dos nós da rede de comportamentos[9].

[9] ver seção 4.9.

Eles são responsáveis por adicionar para onde ir, ou por qual velocidade seguir, se se deve evitar um obstáculo ou fazer a recarga das baterias.

Esses codelets ficam ativos e executáveis conforme a dinâmica da rede de comportamentos. Apesar da rede de comportamentos utilizada não escolher mais de um comportamento por rodada, a dinâmica do ciclo cognitivo e o paralelismo entre seus passos pode levar a situações de mais de um codelet de comportamento ativo em um determinado instante[10].

Na versão atual do agente, não é possível adicionar codelets de comportamento ou adicionar nós à rede de comportamentos em tempo de execução, ou seja, não há aprendizagem de novos comportamentos.

Codelets Motores

Os codelets motores completam os codelets de comportamento por serem responsáveis pela parte do *"como fazer"*, através da ação sobre os atuadores, aquilo que foi determinado pelos codelets de comportamento. Esses codelets se mantêm ativos durante toda a execução do agente.

5.3.3 Memória de Trabalho

A memória de trabalho consiste em um conjunto de registradores que são responsáveis por manter informações temporárias. A maior parte da memória de trabalho está relacionada ao status atual da criatura virtual. O codelet de comunicação constantemente sobre-escreve registradores como velocidade, ângulo da roda, posição e informações sensoriais.

Esse módulo também funciona como uma interface entre os diversos componentes de CAV, por exemplo, entre a consciência e a rede de comportamentos. Alguns codelets, incluindo codelets de atenção, ficam observando o que é escrito na memória de trabalho, a fim de trazer situações relevantes, urgentes ou insistentes. Quando algo é encontrado, esses codelets buscam ter acesso à consciência para informar o sistema da situação. Dessa maneira, as informações importantes da memória de trabalho são difundidas aos demais codelets e a rede de comportamentos (ver figura 5.6).

5.3.4 Mecanismo de Consciência

O mecanismo de consciência, formado pelo *controlador de coalizões*, *controlador de foco de luz*, *controlador de broadcast* e pelo *campo de jogo*, direciona o agente para atender os eventos mais importantes. Os codelets que desejam disputar a consciência entram em "campo" (ou no palco, pela metáfora do teatro de Baars). O codelet mais relevante é selecionado pelo *controlador de foco de luz*, um mecanismo de controle de atenção e ganha o direito de realizar a difusão dos seus resultados. A difusão dessas informações

[10] ver seção 5.3.7, a seguir.

é realizada para todos os codelets cadastrados[11] no *gerenciador de broadcast*.

Como discutido no capítulo 2, esse módulo tem uma função integradora pois permite que as várias partes do sistema sejam avisadas de uma determinada situação ou acontecimento e possam, de maneira colaborativa, chegar a uma solução. O mecanismo de consciência pode ser comparado a um jornal televisivo de grande audiência: muitas pessoas não tem conhecimento de uma determinada situação até que seja transmitido através do jornal. Assim, não é incomum que em algumas situações, como em catástrofes, várias delas tomem uma atitude em relação a isso, enviando mantimentos, dinheiro ou se deslocando para auxiliar as vítimas.

Assim como em um jornal televisivo (sério) em que as notícias mais importantes são veiculadas, o mecanismo de consciência divulga as informações vindas dos codelets de atenção que são mais relevantes de maneira global. Essa é a principal diferença entre esse mecanismo e os tradicionais modelos de quadro-negro em que várias fontes de conhecimento podem ser compartilhadas (Nii, 1986). Por meio do mecanismo de consciência, informações menos relevantes são intencionalmente preteridas, permitindo uma contextualização melhor da situação para o agente e restringindo o escopo de atuação aos codelets mais relevantes para a situação corrente.

A seguir, são detalhados os principais componentes do mecanismo de consciência. Os diversos algoritmos foram retirados de (Bogner, 1999, cap. 5).

Campo de Jogo

O campo de jogo, inspirado na metáfora da arena de (Jackson, 1987), provê aos codelets acesso a consciência. Os codelets que estão interessados em entrar na consciência entram em campo logo antes de executar suas tarefas. Os codelets deixam o campo após terem completado suas atividades e terem ganho o acesso à consciência[12]. Os codelets que entram no campo são armazenados em uma lista.

Gerenciador de Coalizões

O gerenciador de coalizões lê a lista de codelets no campo de jogo, buscando formar coalizões baseadas nos links entre os codelets. Na implementação de CAV o gerenciador de coalizões não é uma *thread* a parte (como é implementado originalmente em (Bogner, 1999)), sendo chamado uma vez a cada ciclo cognitivo (ver seção 5.3.7). A cada rodada o gerenciador de coalizões executa os seguintes passos:

1. Cria uma nova tabela temporária de coalizões;

2. Lê a lista de codelets em campo. Cada codelet encontrado é inicialmente colocado na sua própria coalizão, formando coalizões de um único codelet;

[11] Essa é uma diferença sutil, mas importante se comparada com as implementações de *IDA* e *CTS*. Nesse trabalho, nem todos os codelets recebem o broadcast. Aqueles que nunca se interessariam pelo conteúdo do broadcast consciente não são cadastrados no gerenciador de broadcast. Isso dá um ganho computacional por limitar o escopo do broadcast aos codelets que possivelmente irão utilizá-lo.

[12] A maneira como os codelets entram e saem do campo é um ponto crítico para a questão da formação de coalizões.

3. Verifica todas as associações presentes na coalizão, ou seja, analisa as várias associações de cada codelet presente na coalizão. Para cada associação, o gerenciador de coalizões busca por codelets associados que estão em campo, mas que não pertencem à coalizão. Se o codelet buscado é encontrado, ele é adicionado à coalizão;

4. Sobrescreve a tabela de coalizões com a nova tabela temporária de coalizões;

5. Atualiza todas as associações de codelets. Especificamente, se um codelet está em campo e não está associado a outro codelet também em campo, uma associação é criada. Se a associação já existe, ela é reforçada por uma quantia pequena, assumindo que essa associação não esteja com a força em seu valor máximo[13].

Controlador de Foco de Luz

O controlador de foco de luz determina o conteúdo da consciência. Na metáfora do teatro da teoria do *workspace* global, ele é o foco de luz que brilha sobre uma coalizão de codelets. O controlador de foco de luz é invocado uma vez a cada rodada do ciclo cognitivo e executa os seguintes passos:

1. Calcula a média do nível de ativação dos codelets em cada uma das coalizões do gerenciador de coalizões. A média da ativação é utilizada (ao invés da soma da ativação) para garantir que coalizões com um maior número de codelets não tenham vantagens no acesso à consciência;

2. Para a "coalizão" com o maior nível de ativação, é observado se esse nível é maior que o limiar para entrar na consciência:

   (a) Se for, o controlador de foco de luz seleciona a coalizão para ser a coalizão consciente;

   (b) Caso contrário, é reduzido o limiar para ter acesso à consciência em 10%, e o método é finalizado[14];

3. As associações dos codelets são atualizadas. Tanto as associações entre os codelets que estão juntos em campo, como as dos codelets da coalizão vencedora são reforçadas (ou criadas caso não existam). Entretanto, as associações dos codelets que acessam à consciência juntos tem um reforço significativamente maior (diferença de uma ordem de grandeza);

4. A coalizão vencedora é passada ao gerenciador de broadcast;

[13] Bogner (1999, p.80) afirma que as atualizações da força das associações no campo de jogo, inspirada na teoria do Pandemônio, potencialmente, permite uma evolução do comportamento do sistema durante o tempo. Porém, na implementação de CAV, essas atualizações não produziram esse efeito. Isso aconteceu, provavelmente, pelo baixo número de codelets em campo e, principalmente, por eles não manterem uma regularidade da sua presença simultânea em campo. Por isso, associações geradas são rapidamente desfeitas, uma vez que dois codelets não aparecem juntos com uma frequência significativa para que a associação se perpetue.

[14] Essa redução é inspirada pelas redes de comportamento de Maes (Bogner, 1999, p. 81).

5. Cada codelet que teve acesso à consciência tem sua variável booleana que indica se ele foi escolhido para ir à consciência colocada em *verdadeiro*[15];

6. A coalizão consciente é colocada como *null* para indicar inexistência de coalizão consciente;

Gerenciador de *Broadcast*

O gerenciador de broadcast é responsável por manter o registro de todos os codelets que desejam receber as informações da coalizão vencedora. Ele dissemina a informação consciente através da sequência de passos a seguir:

1. Varre a coalizão consciente, pegando as informações de cada codelet que serão divulgadas. Todos as informações são colocadas em uma única *hashtable*;

2. Insere um *timestamp* na *hashtable*;

3. Realiza o broadcast propriamente dito a todos os codelets cadastrados[16].

5.3.5 Formação de Coalizões

Em IDA, os codelets de atenção são fundamentais na formação de coalizões que são tipicamente formadas por codelets de atenção e codelets de informação (Franklin, 2005). Esses últimos, ou já possuiam *links* com esses codelets de atenção (adicionados no projeto do sistema), ou têm *links* criados durante a disputa pela consciência. Por meio desses *links*, os codelets de atenção podem recrutar (instanciar) codelets de informação relevantes para reforçar uma coalizão que possivelmente irá atender melhor a situação corrente.

Na implementação de CAV, optou-se por remover os codelets de informação pois todas as informações estão disponíveis na memória de trabalho (em IDA as informações poderiam não estar completas pela falta de dados em um email, por exemplo). Os codelets de comportamento também foram removidos da disputa pela consciência. Os resultados da execução de um codelet de comportamento (alteração nas proposicões do estado da rede de comportamentos) são monitorados também por codelets de atenção ou codelets de expectativa. Isso reduziu substancialmente o número de codelets na disputa pela consciência e removeu a existência de coalizão na implementação atual de CAV.

5.3.6 Rede de Comportamentos

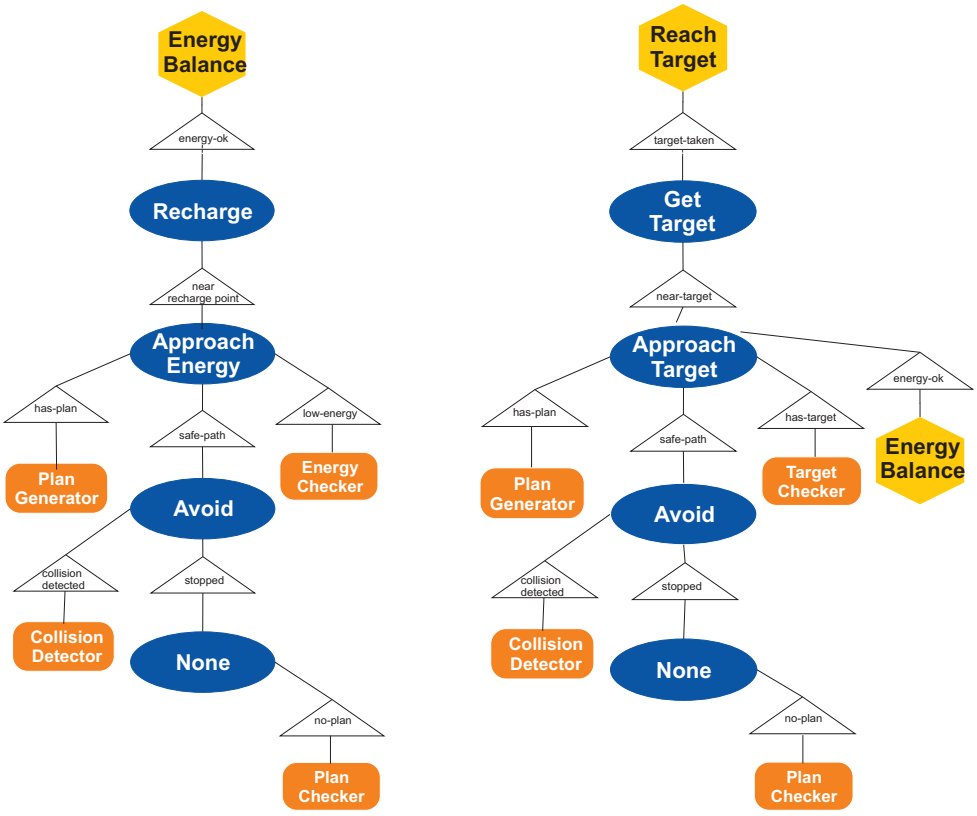
A rede de comportamentos serve como uma estrutura de decisão e um mecanismo de planejamento implícito. As decisões realizadas pela rede de comportamentos são

[15] Por essa variável os codelets podem decidir por continuar em campo, caso não consiga obter acesso à consciência da primeira vez.

[16] Originalmente o algoritmo de (Bogner, 1999) armazena um certo número de *brodcasts* em uma memória de curto prazo. Isso não é realizado em CAV.

feitas inconscientemente (ou seja, sem que o comportamento seja explicitamente publicado pela consciência). Entretanto, a escolha do comportamento executável é influenciada pela consciência. Além disso, ao ser executado o comportamento normalmente gera novas chamadas à consciência (novos *broadcasts*). Essa intercalação entre *consciente-inconsciente-consciente* é chamada de *seleção de ações mediada pela consciência* (Negatu, 2006, p. 105).

Nesse trabalho, foi utilizado para o desenvolvimento da rede de comportamentos um framework implementado por Sidney D'Melo, pertencente ao grupo de Stan Franklin. Esse framework é uma implementação simplificada de (Negatu, 2006). Utilizando o trabalho de D'Melo foi possível projetar uma rede de comportamentos através de um arquivo XML. A figura 5.7 mostra um exemplo das estruturas básicas da construção da rede.



(a) Manter o balanço energético (b) Alcançar a meta

Figura 5.7: Estruturas da rede de comportamentos. As elipses representam os nós de comportamento, os triângulos, as proposições adicionadas ao estado da rede, o pentágono representa o objetivo e os retângulos são os codelets de atenção que interferem no estado da rede. A figura 5.7(a) mostra um ramo da rede para resolução do objetivo *"alcançar a meta"* e a 5.7(b) o ramo para o objetivo *"manter o equilíbrio energético"*

(Negatu, 2006) divide a rede de comportamentos em cadeias de comportamento, a fim de não precisar tratar a rede inteira o tempo todo, o que pode ser conveniente quando a rede é muito grande. Em IDA os codelets de comportamento estão sempre

ativos, esperando por broadcasts e quando eles se avaliam relevantes para a situação eles instanciam a cadeia de comportamentos da qual fazem parte. No *CAV* todas as cadeias são carregadas na inicialização do sistema e permanecem durante a execução do agente. *CAV* também se diferencia no modo de execução dos comportamentos, permitindo sobreposição entre alguns deles. Isso permite que o comportamento *"avoid"*, responsável por tratar colisões, seja habilitado mesmo quando ainda não foi terminado o comportamento de *"approach target"*, que é responsável por levar o veículo até a meta. Entretanto, caso a rede escolha *"approach target"*, enquanto ainda há codelets desse comportamento ativo, então os codelets do comportamento escolhido não são novamente ativados.

Para resolver esse problema de comportamentos concorrentes, Dorer (1999) desenvolveu uma melhoria que adiciona à rede de Maes a possibilidade de comportamentos concorrentes selecionando-os não somente pela energia de ativação mas pelos atuadores que são influenciados pelo comportamento. Assim, é possível ter vários comportamentos concorrentes desde que eles não concorram pelo mesmo atuador. Essa abordagem foi utilizada com sucesso como mecanismo de seleção de ação para agentes em jogos de computador em (Pinto, 2005) e poderia ser uma alternativa à rede de comportamentos de Negatu.

5.3.7 O ciclo cognitivo de CAV

O ciclo cognitivo de CAV (CCC) é inspirado no *ciclo cognitivo de IDA*[17]. As maiores alterações estão na remoção dos três primeiros passos originais: *percepção* (relativo à interpretação das entradas sensoriais), *percepto no buffer pré-consciente* (responsável por armazenar o percepto na memória de trabalho), associações locais (relativo ao armazenamento e recuperação de associações locais na memória episódica transiente (TEM) e da memória associativa de longo prazo (LTM). Essa última alteração é bastante óbvia uma vez que CAV não implementa TEM ou LTM. Nos outros dois primeiros casos, a remoção está ligada ao domínio do problema. CAV não processa cadeias de caracteres como IDA ou CTS. Portanto, CAV não utiliza uma Slipnet; as informações vindas do simulador são capturadas em um ciclo próprio do codelet de comunicação, paralelo ao CCC. Por último, uma alteração significativa está na remoção do passo 6 (recrutamento de recursos), pois a "resposta" dos codelets registrados no gerenciador de *broadcast* também ocorre em paralelo ao ciclo de CAV, não no seu interior. Além disso, como não se optou pela utilização de codelets de informação, esses não têm como ser recrutados.

Assim, o CCC possui cinco passos explicados abaixo[18]:

(1) Competição pela consciência

Durante todo o tempo de execução do agente, os codelets procuram por informações ou estados que sejam relevantes para eles e que lhes permitam entrar em execução. Em especial, os codelets de atenção ficam constantemente avaliando os sensores e a

[17] ver seção 4.10.

[18] Adaptado de (Baars & Franklin, 2003). As partes do ciclo que são comuns a IDA e CAV são colocas em *itálico*.

rede de comportamentos para levar à consciência os eventos urgentes, relevantes ou insistentes. Essa competição é feita pelo "controlador de foco de luz", que recebe as coalizões formadas pelo "controlador de coalizões".

A competição é realizada uma única vez a cada ciclo cognitivo. Originalmente ConAG implementa os *controladores de foco de luz e de coalizões* utilizando *threads* separadas para cada um deles. Essas *threads* tinham originalmente um ciclo próprio da ordem de segundos[19], uma vez que a aplicação - CMattie (Song, 1998) - não necessitava de processamento em tempo real. Como um ciclo longo desses controladores inviabiliza a aplicação, em CAV foi removida a implementação do controlador de coalizão e do controlador do foco de luz em *thread* própria, adicionando-se uma chamada por ciclo a cada controlador.

(2) Broadcast Consciente

A coalizão vencedora no passo anterior ganha o direito de ter seu conteúdo propagado a todos os codelets assinantes do gerenciador de broadcast. Uma grande diferença de implementação para *IDA* é que no *CAV* os codelets de comportamento, expectativa e motor não são cadastrados no gerenciador de *broadcast*. Esses codelets não participam da competição pela consciência e também não têm seus resultados propagados diretamente por eles. No caso dos codelets de comportamento, que podem produzir novas proposições no estado da rede de comportamentos, eles têm seu resultado levado para a consciência por determinados codelets de atenção que procuram por proposições específicas. Na prática isso leva à divulgação de resultados relevantes de determinados comportamentos para todos os codelets registrados, quando esses codelets de atenção atingem a consciência.

(3) Configurando a hierarquia de contextos de objetivos

Nesse estágio, CAV atualiza a rede de comportamentos com todas as proposições novas que foram geradas desde a última atualização e incorpora informações novas e mais precisas. Os objetivos são verificados e atualizados. Nesse passo, também é possível adicionar ou remover objetivos, ou alterar o seu valor associado, de acordo com a situação corrente.

(4)Escolha da ação

Após ter todo o cenário montado pela rodada da competição de consciência e *broadcast* do conteúdo dos codelets da coalizão vencedora e pelas atualizações realizadas no passo anterior, a rede de comportamentos escolhe um comportamento para ser executado. Quando um comportamento não é encontrado, a rede é chamada novamente para uma nova rodada até o limite de 100 rodadas.

(5) Realização da ação

Os codelets associados ao comportamento escolhido são executados e as atividades especializadas são inicializadas. Diferente de IDA e CTS, nem todo comportamento tem codelets de expectativa associado. Outra particularidade importante é que os

[19] O ciclo de cada codelet dura na ordem de centenas de milisegundos.

comportamentos têm uma certa duração no tempo, como por exemplo se mover do ponto A ao ponto B. Após a inicialização dos codelets de comportamento, é dado início ao outro ciclo sem que seja esperado o fim da execução dos codelets de comportamento. Isso aumenta a sobreposição de ciclos cognitivos e permite que seja iniciado o processamento de problemas que possam vir a acontecer, como uma colisão ou a detecção de falta de combustível.

5.3.8 Sistema de Controle

Essa seção descreve com mais detalhes os algoritmos de controle que foram utilizados na construção de CAV.

A entrada do controlador é dada pela ênupla $(x, y, \psi, f, v, \theta, \rho, \phi, c, s)$, onde:

x, y, ψ correspondem à posição do veículo;

f corresponde à carga das baterias do veículo (em %);

v corresponde à velocidade nominal;

θ corresponde ao ângulo das rodas;

ρ corresponde à distancia do centro do veículo ao sensor remoto;

ϕ corresponde ao ângulo entre o eixo longitudinal do veículo e o sensor remoto;

c corresponde aos 4 sensores de contato (c_1, c_2, c_3, c_4);

s corresponde à matriz de 49 sensores de informação remota.

A saída do controlador corresponde à seguinte ênupla: (v, θ, ρ, ϕ), onde:

v corresponde à velocidade nominal a ser aplicada;

θ corresponde ao ângulo na roda a ser aplicado;

ρ corresponde à distância do sensor remoto, a ser aplicada;

ϕ corresponde ao ângulo do sensor remoto, a ser aplicado.

A ênupla de entrada, armazenada na memória de trabalho, é atualizada constantemente por um codelet de comunicação, chamado *"Status Reader"*. Os codelets de atenção observam os parâmetros da entrada e caso alguma situação relevante seja percebida, eles buscam ter acesso à consciência. Por exemplo, o codelet *"Energy Checker"* verifica o nível de energia das baterias e, caso haja pouca energia, ele busca acessar à consciência para anunciar aos demais codelets e alterar os parâmetros da rede de comportamentos, com a finalidade de iniciar um tratamento adequado à situação.

A memória de trabalho contém um modelo de mundo, que está vazio no início da simulação e é construído conforme a interação da criatura com o ambiente. A percepção dos obstáculos é gerada pelo codelet de percepção *"Obstacle Recorder"*. Esse

codelet, ao encontrar algum conteúdo nos sensores s, realiza uma varredura na matriz de sensores, da "esquerda para a direita" e "de cima para baixo" para reconhecer os obstáculos (ver figura 5.8). O algoritmo de reconhecimento utiliza uma heurística bastante simples: iniciando do canto esquerdo superior, esse codelet tenta encontrar, em uma linha, os sensores ativos contínuos e de mesma cor. Para uma dada linha, após determinada as extremidades inicial e final de sensores ativos, o algoritmo busca ampliar o objetos percebido, procurando, nas linhas inferiores, sensores ativados com as mesmas características.

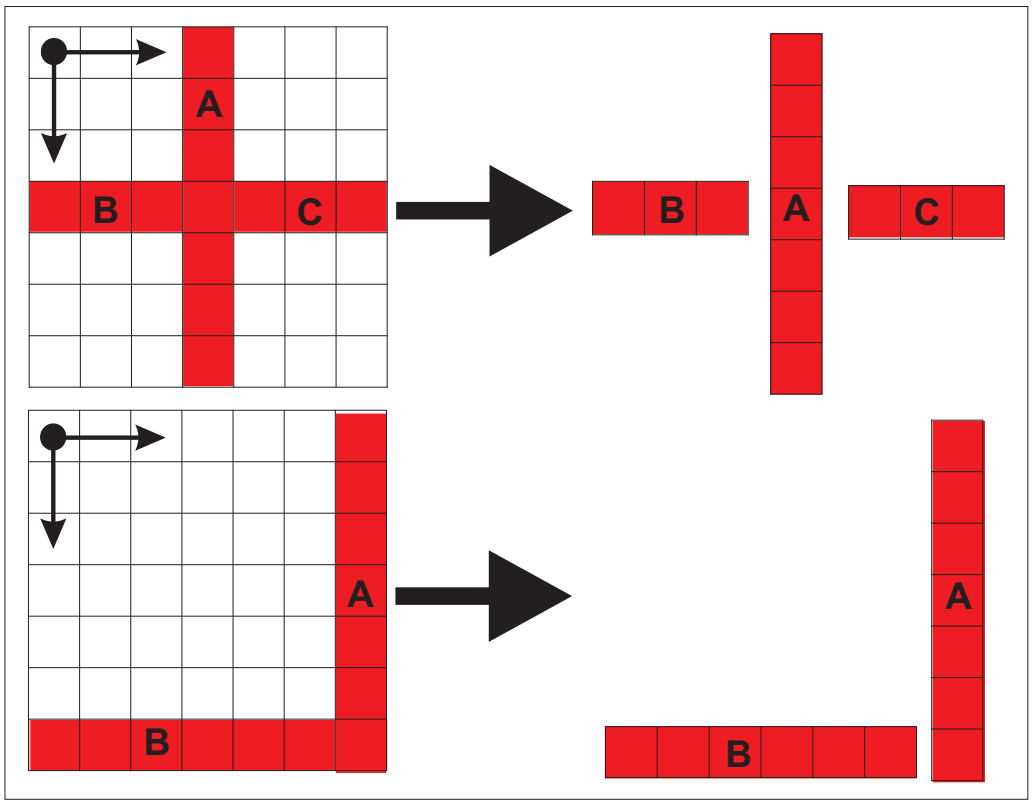


Figura 5.8: Reconhecimento de padrões. À esquerda é colocado dois exemplo de enquadramento hipotético de padrões. À direita são mostrados os padrões reconhecidos, resultantes do uso do algoritmo de reconhecimento.

Após ter sido realizado o reconhecimento dos padrões, o codelet *Obstacle Recorder* verifica procura adicionar ao modelo de mundo. Nessa fase, são utilizadas três regras de integração de objetos (ver figura 5.9), como sugerido em (Gudwin, 1996, p. 113):

> A primeira diz que se dois objetos estão alinhados horizontalmente, eles na verdade correspondem a um único objeto com a dimensão correspondente à sobreposição dos outros dois. A segunda regra diz que se dois objetos estão alinhados verticalmente, então eles correspondem a um único objeto. A terceira regra diz que se um determinado objeto está integralmente contido em outro objeto, ele na verdade é apenas uma parte desse, sendo descartado.

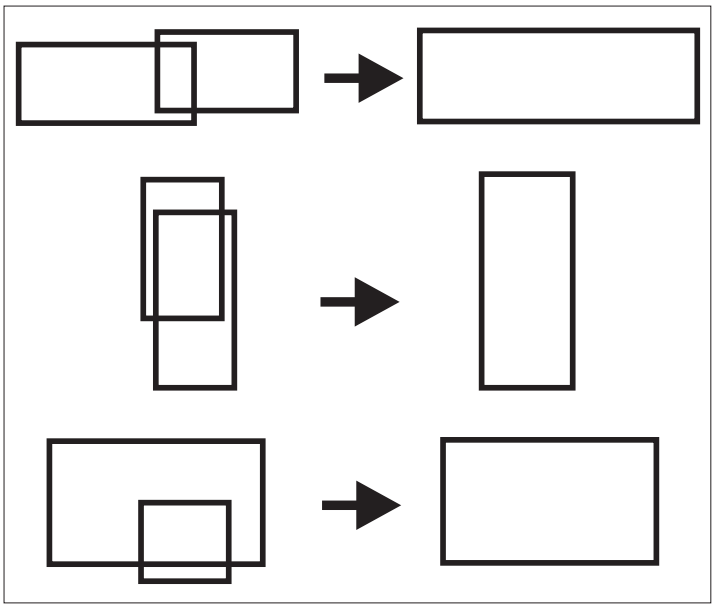


Figura 5.9: Regras de integração de objetos.

Nos experimentos realizados, os objetos podem ser obstáculos ou pontos de recarga de energia. Os obstáculos são armazenados adicionando uma margem de segurança (obstáculo extendido), que indica à criatura que ela está muito próxima do obstáculo. Os pontos de recarga são adicionados sem essa margem, uma vez que a criatura pode passar sobre esses objetos sem ocorrer colisão. Quando um novo obstáculo é adicionado ao mundo, o codelet *"Ladmark Handler"* adiciona *landmarks* ao redor do obstáculo (ver figura 5.10), que servirão de guias para o algoritmo de planejamento.

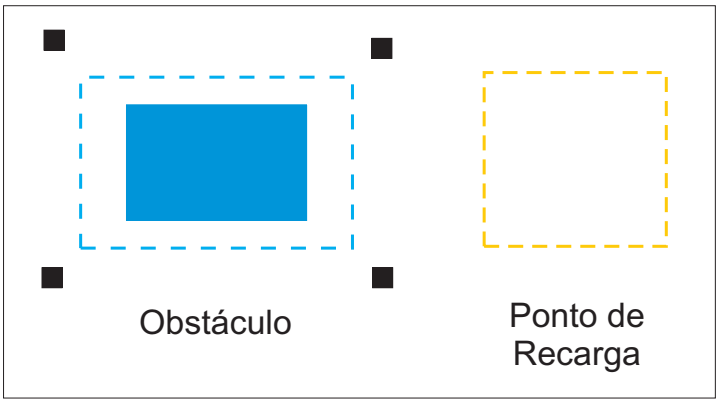


Figura 5.10: Modelagem dos obstáculos no modelo de mundo.

Uma meta (ponto para onde a criatura deve se dirigir) é gerada aleatoriamente, tendo o cuidado para que ela não seja inviável, por exemplo, posicionada dentro de um obstáculo. A meta, os *landmarks* e a posição da criatura formam o conjunto de vértices do grafo, cujas as arestas são todas as ligações entre esses vértices que não cruzam algum obstáculo (ver figura 5.11). Utilizando esse grafo, o codelet de atenção *"Plan*

Generator" cria um plano formado pelo caminho mínimo entre a posição da criatura e a meta, por meio do algoritmo de Dijkstra, toda vez que for detectado que não há um plano.

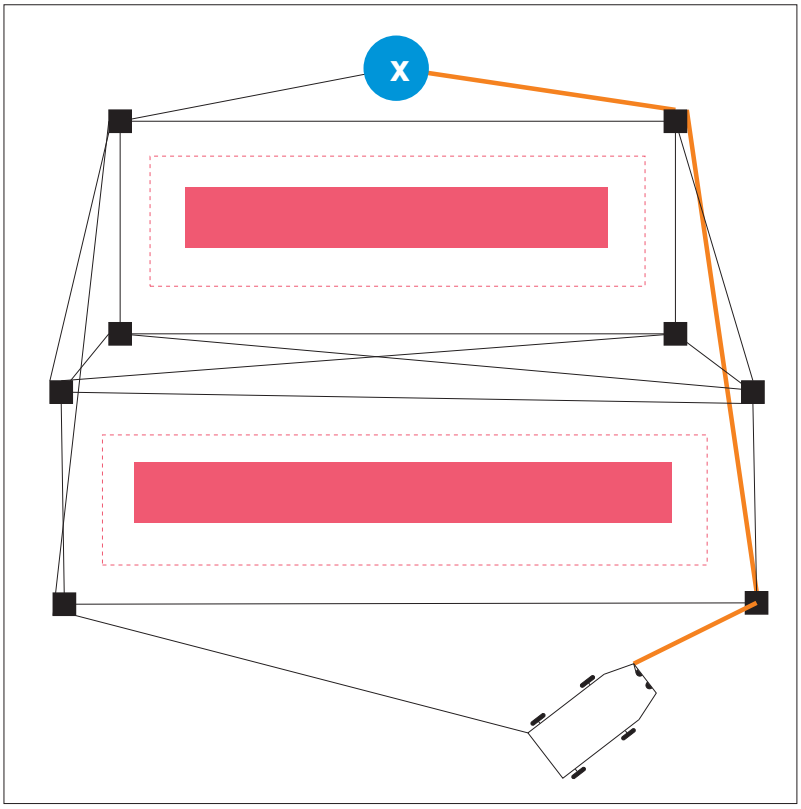


Figura 5.11: Planejamento de CAV.

5.3.9 Características interessantes de CAV

A proposta de *CAV* traz algumas características interessantes para a resolução do problema exemplo.

Arquitetura pouco acoplada

A implementação do núcleo do agente com a rede de comportamentos e da consciência, ligados através do ciclo cognitivo e pelas mensagens entre os codelets via *broadcast* do mecanismo de consciência, cria uma arquitetura distribuída pouco acoplada. Essa característica de baixo acoplamento, sistema paralelo (ver figura 5.13), permite de maneira relativamente simples a inserção de novos codelets, a fim de adicionar no sistema novas funcionalidades[20]. A estrutura montada permite também a adição de novos módulos (como os de memória de longo prazo e memória episódica), que venham a ser desenvolvidos posteriormente.

[20]O maior desafio do aumento de codelets está nas dificuldades inerentes à programação paralela e do gerenciamento das threads.

Processamento Paralelo e Serial

CAV consegue agrupar o processamento paralelo dos codelets e o funcionamento serial do mecanismo de consciência. Essa estrutura gera uma fonte de processamento paralelo rápido e eficiente. A parte serial (simbolizada pela consciência) é uma maneira de priorizar os resultados dos processamentos paralelos, de modo a atender os eventos mais relevantes primeiro.

Essa simbiose, entre o mecanismo serial de consciência e o paralelismo dos codelets, está de acordo com a hipótese defendida por Dennett: "A consciência humana (...) pode ser melhor entendida como a operação de uma 'máquina virtual de von Neumman' *implementada* em uma arquitetura paralela do cérebro" (Dennett, 1991, p. 210). Dessa maneira, a ABF poderia ser pensada como uma instância de tal hipótese, considerando-se a saída serial do mecanismo de consciência como a saída de uma *"máquina de von Neumman"* e a infraestrutura de codelets, a *"arquitetura paralela do cérebro"*.

Decisões sobre dados atualizados

Devido à atualização constante da memória de trabalho pelo codelet de comunicação, todos os pontos de decisão se referem aos dados mais recentes possíveis obtido pelo agente. Em um controlador em que os dados são obtidos e então um longo processamento é realizado até que um novo dado seja considerado, é muito provável que, ao final, os dados estejam demasiadamente desatualizados. Esse efeito foi minimizado nos trabalhos anteriores (Gudwin, 1996; de Toro, 2007), pois havia uma sincronização entre os ciclos do controlador e do simulador. O fato de ter os dados mais recentes disponíveis a todo tempo auxilia também os codelets de atenção que podem se excitar e tentar competir pela consciência o quanto antes. Alguns codelets alteram os estados da rede de comportamentos, o que pode levá-la a escolher um comportamento mais conveniente para a situação.

Sumário Executivo

Além dos dados estarem sempre atualizados no momento da tomada de decisão, as informações transmitidas pelo mecanismo de consciência podem ser vistas como um sumário executivo da percepção, o qual destaca o que há de mais importante a ser tratado naquele momento (ver figura 5.15) dentro de todas as informações captadas nos diversos sensores do agente. Nesse sentido, a análise feita pelo *Gerenciador de Foco de Luz* é baseada em uma visão holística da situação e não apenas no tratamento pontual de um ou outro problema que esteja ocorrendo.

5.3.10 Análise quantitativa

Para ilustrar quantitativamente a análise realizada, foram coletados alguns dados durante uma simulação, considerando o mundo virtual modelado o da figura 5.4. Foram armazenados o número de *threads* ativas, o número de codelets no campo de jogo (tentando acessar à consciência) e o codelet consciente (aquele que de fato ganhou o direito de fazer o *broadcast*). Os dados do primeiro minuto de simulação são mostrados

nas figuras 5.13, 5.14, 5.15. O experimento foi realizado colocando-se o simulador em uma máquina e o controlador em outra (como mostrado na figura 5.5). O simulador e o controlador foram executados em máquinas Quadricore Intel.

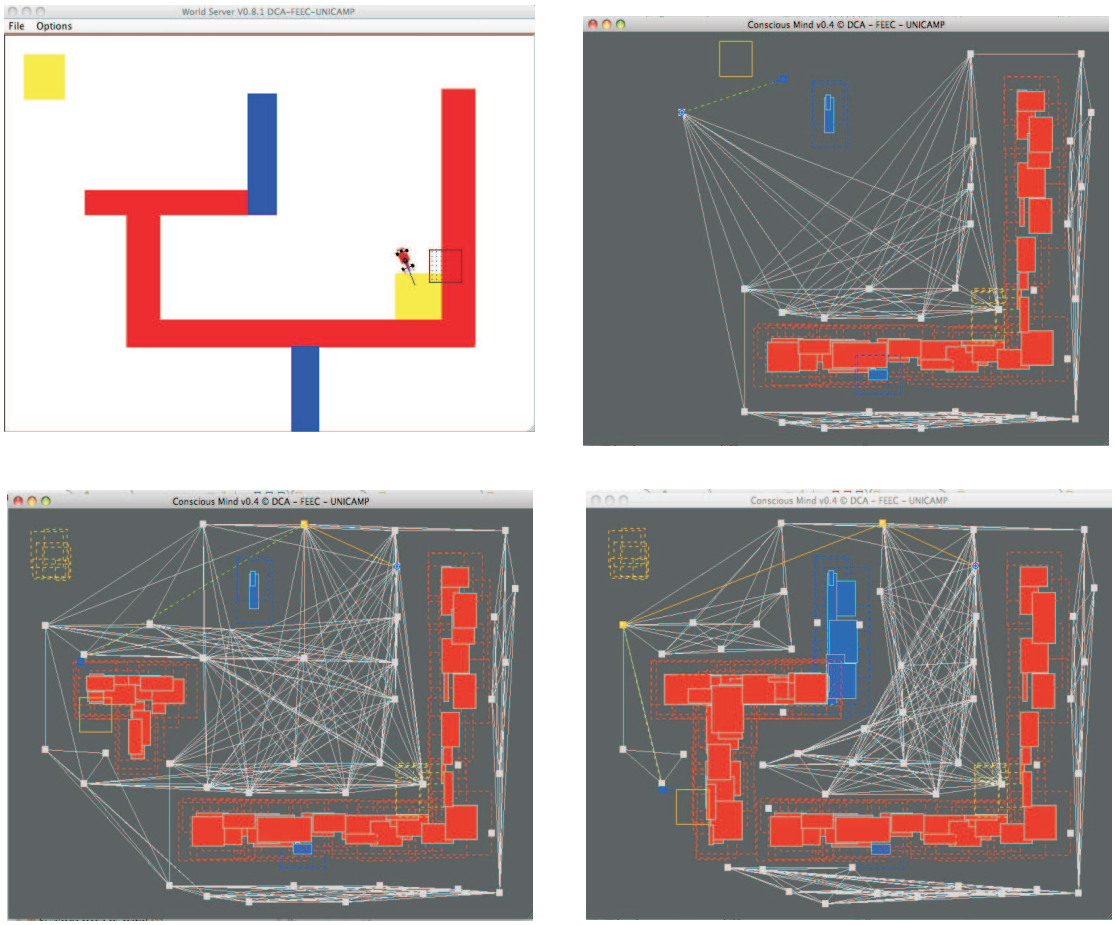


Figura 5.12: *Screenshots* das simulações.

A figura 5.13, mostra o número de *threads* ativas devido ao processamento dos algoritmos de controle. São excluídas desse gráfico qualquer outra *thread*, como as relacionadas às atualizações gráficas da janela do programa. O número de *threads* ativas é mantido estável em uma faixa de 6 a 11 *threads*, o que mostra que não há explosão combinatorial.

A figura 5.14, traz um gráfico com o número de codelets ativos simultaneamente. O número é bastante reduzido devido à decisão de não implementar os codelets de informação e da remoção dos codelets de comportamento da disputa pela consciência. A velocidade de processamento da máquina também auxilia na diminuição do número de codelets em campo simultaneamente. Esse número reduzido prejudica a formação de associações (*links*) entre os codelets, dificultando o aparecimento de coalizões que poderiam ser geradas pela criação de associações entre os codelets[21]. A adição de novos

[21] Passo 5 do algoritmo do gerenciador de coalizões, seção 5.3.4.

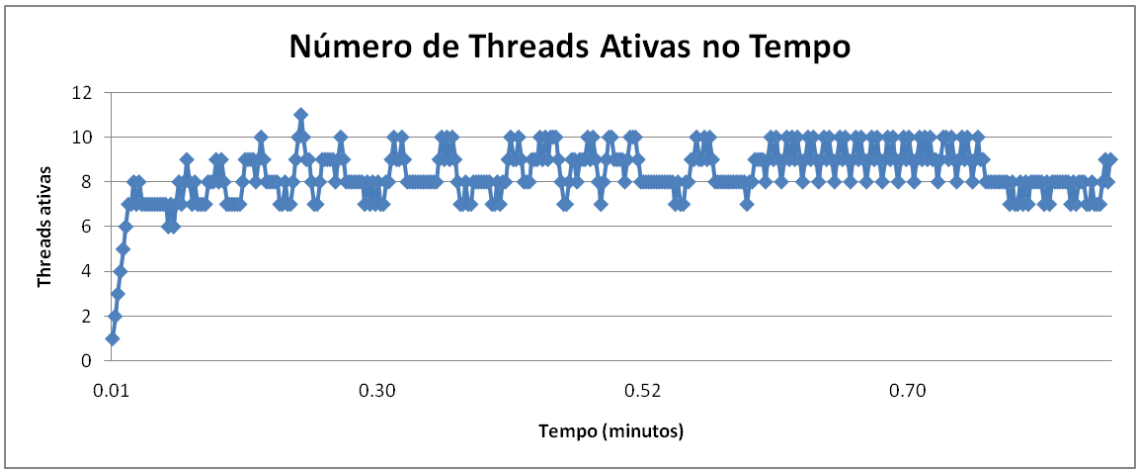


Figura 5.13: Número de *threads* ativas. São consideradas apenas as *threads* referentes ao algoritmo de controle. É importante ressaltar que o número de *threads* se mantém controlado, dentro de uma faixa de 6 a 11 *threads*.

mecanismos de percepção ou a adição do processamento de sensores reais deve aumentar a concorrência e se beneficiar mais do uso do mecanismo de consciência.

Por último, a figura 5.15 mostra, para um minuto de simulação, os codelets que ganham acesso à consciência. Esses codelets influenciam o fluxo de controle após o *broadcast* das informações de seus processamentos.

5.3.11 Comentários sobre a implementação

Nesse projeto, foi utilizada a linguagem Java, principalmente devido ao fato de esta ser a tecnologia utilizada tanto em *IDA* como nos trabalhos anteriores (Suárez, 2000; de Toro *et al.* , 2007). A tecnologia Java apresentou um bom suporte a múltiplas linhas de execução através das Threads e ao mecanismo de *"publish-subscribe"* através dos *Listeners*. Diferente de *CTS*, em que a arquitetura foi reescrita do zero, nesse trabalho foram reutilizados os módulos de Rede de Comportamentos e de Consciência da tecnologia *IDA*, com adaptações e melhorias.

Dubois (2007b) afirma ter tido problemas com as várias mudanças ocorridas em *IDA* durante o desenvolvimento de seu agente *CTS*. Nesse trabalho, foi iniciada a implementação, tendo como base apenas a bibliografia disponível, mas, após firmado o acordo com a Universidade de Memphis, passou-se a utilizar parte da tecnologia *IDA*. As maiores dificuldades foram agrupar os módulos para trabalhar em conjunto, uma vez que não foi enviado um conjunto coeso da tecnologia. Contudo, os maiores problemas se concentram na montagem da arquitetura do agente pois o projeto da arquitetura é extremamente dependente dos codelets criados pelo projetista. Isso requer um projetista especialista no problema e que tenha conhecimentos do funcionamento da ABF. Além disso, a pouca formalização contribui para muitas dúvidas que precisam ser resolvidas durante a fase de desenvolvimento.

Um ponto que também representa problemas nesse trabalho, assim como no CTS, é a questão de escalonamento das múltiplas *threads*. Os testes realizados com *CAV* mos-

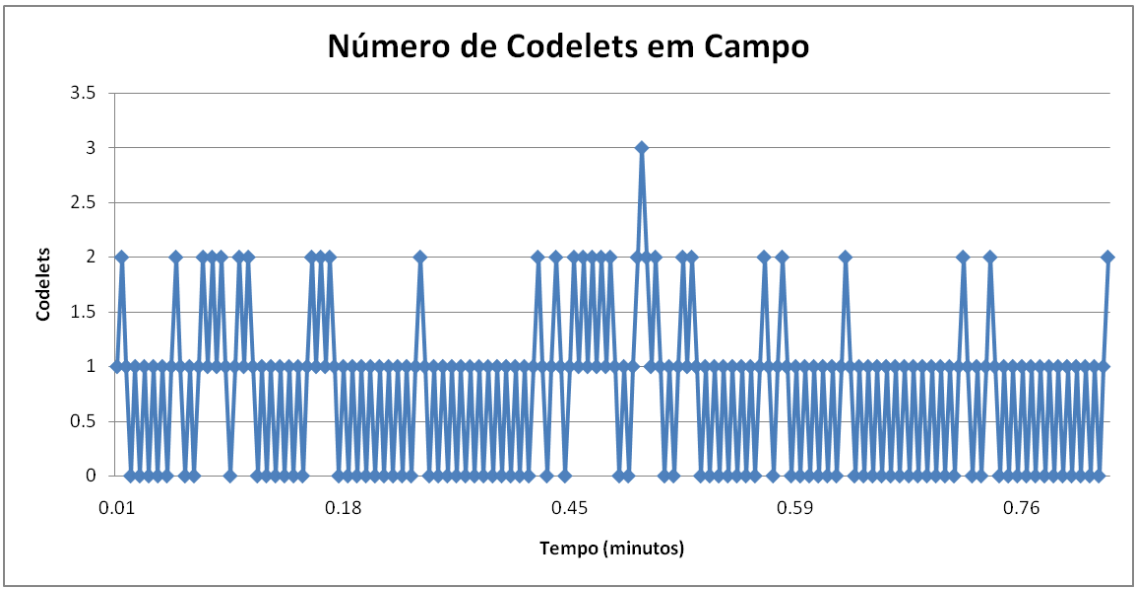


Figura 5.14: Número de codelets no campo de jogo simultaneamente.

traram que algumas vezes as *threads* sofrem atrasos não relacionados com a arquitetura em si. Dubois (2007b, p. 107) afirma que esses problemas com *threads* podem ser minimizados com a implementação de um escalonador próprio, por essa solução oferecer um maior controle sobre a ordenação e execução das threads.

Para concluir, durante a implementação, houve questões de *debugging* relacionadas com a programação paralela. Em geral não é possível utilizar um debugger convencional para o acompanhamento passo-a-passo de uma *thread* uma vez que isso limita o funcionamento do programa e reduz o funcionamento paralelo das *threads*. Assim, a remoção de erros foi realizada através da análise de logs que eram gerados pelo framework Log4J.

Arquitetura de software

A arquitetura de software é bastante modularizada (ver figura 5.16). Os principais módulos, ou "pacotes" em Java, são o *cav*, *legacyclient*, *rbnt* e *conag*.

Os pacotes *rbnt* e *conag* foram providos pela Universidade de Memphis, sendo que o primeiro se refere à implementação genérica da rede de comportamentos de Maes e o segundo, à implementação de (Bogner, 1999). Esses pacotes não sofreram atualizações diretamente. Além disso, como pode ser visto na figura 5.16, não há dependência entre o pacote *conag* e os demais. Isso acontece devido à implementação de CAV utilizar apenas as classes principais de *conag* referentes ao mecanismo de consciência, que foram adicionadas ao pacote *consciousness* de *cav*.

O pacote *legacyclient* possui o cliente legado do trabalho de (de Toro, 2007). CAV faz uso de algumas classes de modelagem desse pacote, como o modelo de obstáculos e do próprio veículo. O pacote *cav* é o principal pacote da arquitetura e implementa os diversos módulos de ligação entre os códigos legados e todas as classes referentes ao domínio do problema. *Cav* traz também a implementação da classe *"Teather"*, que tem

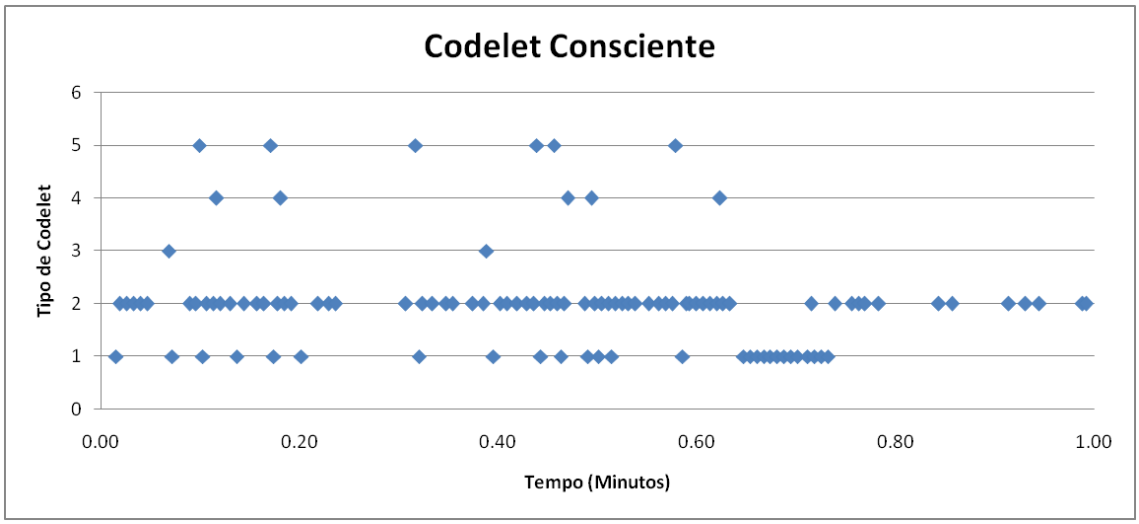


Figura 5.15: Codelet consciente. Nesse gráfico é mostrado o acesso à consciência por um codelet. À cada codelet foi dado um determinado número, sendo 1 - *"Plan Generator"*, 2 - *"Obstacle Recorder"*, 3 - *"Target Carrier"*, 4 - *"Collision Detector"*, 5 - *"Path Checker"* e 6 - *"Energy Checker"*. Para esses dados é possível notar uma dominância nos codelets de planejamento (*Plan Generator*) e de manutenção de obstáculos no modelo de mundo (*Obstacle Recorder*).

.

um papel semelhante à ideia original da sala de um teatro. Essa classe é responsável por criar as instâncias de codelets necessária para a execução do programa e também por gerenciar a implementação do ciclo cognitivo. A implementação da classe *"Theater"* foi inspirada na construção de CTS (Dubois, 2007b, p. 107). *Theater* realiza o ciclo cognitivo, em uma *thread* própria, desempenhando, a cada rodada, os seguintes passos:

1. Chama o gerenciador de coalizões para realizar o cálculo das coalizões em campo;
2. Chama o controlador de foco de luz para fazer o cálculo da coalizão vencedora;
3. Chama o gerenciador da rede de comportamentos para atualizar os estados da rede com as novas proposições e também para atualizar os objetivos;
4. Chama o gerenciador da rede de comportamentos para gerar as rodadas de escolha do comportamento que será executado;
5. Executa os codelets de comportamento e de expectativa do comportamento escolhido quando necessário (caso o comportamento já não esteja sendo executado);
6. Dorme por um determinado tempo (da ordem de centenas de milisegundos).

5.4 Conclusão

Nesse capítulo, foi explicado o ambiente de simulação e o problema exemplo utilizados nesse trabalho. Nele também mostrou-se uma aplicação da ABF no desenvolvimento

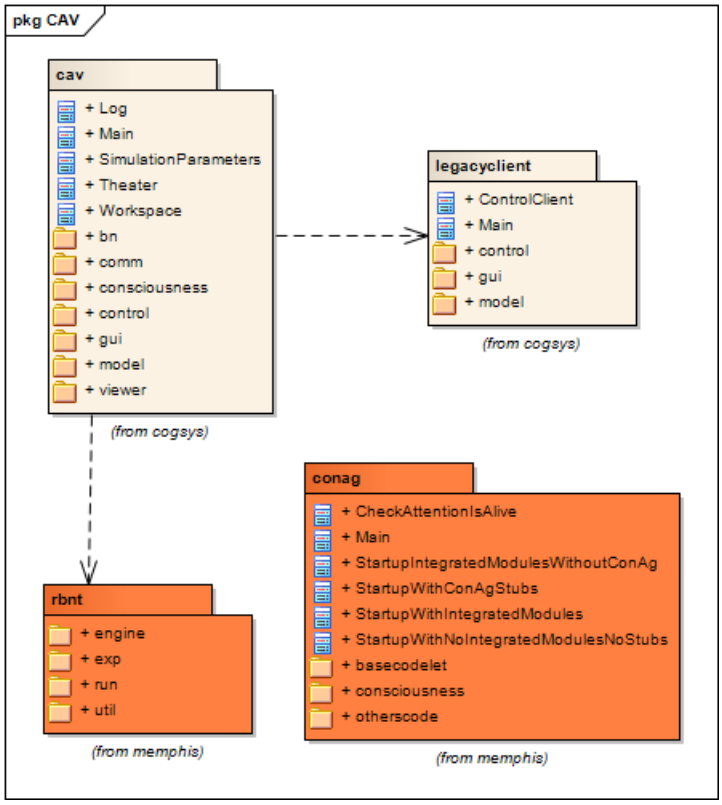


Figura 5.16: Principais pacotes implementação de *CAV*. Os pacotes em laranja foram providos pela Universidade de Memphis.

do controlador de uma criatura virtual em um problema de navegação autônoma. A ABF foi implementada no simulador e suas propriedades foram discutidas.

No próximo capítulo, tem-se a conclusão da dissertação, e os trabalhos futuros.

Capítulo 6

Conclusão e Trabalhos Futuros

> Por uma coisa eu lutaria até o fim, tanto em palavras como em atos se eu pudesse - que se nós acreditássemos que devemos tentar descobrir o que não é sabido, seríamos melhores e mais corajosos e menos preguiçosos do que se acreditássemos que aquilo que não sabemos é impossível de ser descoberto e que não precisamos nem mesmo tentar.
>
> Sócrates, em *Mênon*, de Platão

6.1 Contribuições

Esse trabalho analisou a arquitetura Baars-Franklin de consciência artificial, colocando-a no contexto de outras implementações de modelos de consciência. A ABF foi aplicada no desenvolvimento de um controlador para uma criatura virtual em problema exemplo de navegação autônoma. O desenvolvimento e testes fizeram uso do ambiente de simulação e modelagem desenvolvido anteriormente pelo grupo de pesquisas em cognição artificial (GRACO) do DCA/FEEC/UNICAMP. Além disso, através de uma parceria entre a Universidade de Memphis e a Universidade Estadual de Campinas, foi franqueado o acesso ao mecanismo de consciência implementado em (Bogner, 1999) e à rede de comportamentos implementada por D'Mello, ambos do grupo do Prof. Stan Franklin. Esses módulos puderam ser modificados e integrados à aplicação realizada nesse trabalho.

As principais contribuições desse trabalho são:

- realização de uma revisão bibliográfica sobre o tema de consciência, ressaltando as principais pesquisas para os estudos de consciência artificial;

- realização de um estudo teórico dos principais modelos de consciência artificial, suas aplicações, com foco em arquiteturas cognitivas computacionais;
- refinamento do ambiente de simulação, tornando-o mais próximo da realidade;
- desenvolvimento de uma nova interpretação da ABF como uma instância do conceito de consciência postulado por Dennett, sendo o mecanismo de consciência responsável pela serialização dos resultados provenientes da computação paralela provida pela estrutura de codelets. Esse fato não havia sido postulado anteriormente pelo grupo de Franklin, constituindo-se uma das contribuições teóricas deste trabalho;
- desenvolvimento de um agente, demonstrando o uso da arquitetura Baars-Franklin e a verificação das vantagens e desvantagens de uso dessa tecnologia;
- execução de um estudo de vanguarda no Brasil que ofereceu um ganho de experiência e aprendizado para os membros do GRACO/DCA/FEEC/UNICAMP, abrindo novos caminhos que podem ser trilhados na pesquisa acadêmica de consciência artificial.

6.2 Limitações

A melhor qualidade da ABF, que permite que ela seja aplicável a uma vasta gama de problemas computacionais, é também a sua maior limitação: seu alto grau de generalidade. Isso, associado com a falta de formalismo da teoria, exige muito do projetista, que não basta ser especialista no problema que irá tratar, mas, também precisa tomar muitas decisões durante as fases de projeto e implementação. Isso demonstra, de certo modo, uma imaturidade da teoria.

A falta de formalização da teoria original também corrobora uma dificuldade na reutilização direta dessa tecnologia para outros problemas, como a ideia de um *framework* para desenvolvimento de agentes conscientes que (Bogner, 1999) visava originalmente. Mesmo a implementação disponível no ConAg não permite o desenvolvimento de agentes conscientes, diante do comprometimento deste *framework* com as aplicações alvo para as quais ele foi concebido. Todos os trabalhos que se propõem a ser uma estrutura genérica para o desenvolvimento de agentes conscientes devem ser olhados com ceticismo, pois um longo caminho ainda deve ser percorrido, tanto do lado teórico, como nas implementações.

Em relação à arquitetura de software de CAV, tanto o ambiente em si, como o próprio agente desenvolvido tratam de comportamentos em um escopo bastante restrito. Esse ambiente pode ser evoluído adicionando-se novos codelets e também outras funcionalidades no ambiente de simulação, como mover, remover ou adicionar objetos ao ambiente.

Outro ponto acerca da implementação atual de CAV é a falta de um mecanismo de aprendizado, memórias (memória associativa de longo prazo e memória episódica), mecanismos de metacognição e emoções. Essa funcionalidades, potencialmente, podem trazer muitos ganhos para o agente, como a geração de novos comportamentos

(codelets), geração de automatismos e aprendizado e uso de resultados em situações anteriores.

Por fim, as teorias de consciência, que estão em constante adaptação e evolução, e a ligação dessas teorias com seus correlatos computacionais, ainda não estão consolidadas. Mesmo com várias páginas de artigos e teses tentando contrapor teorias de consciência à sua contraparte computacional, ainda há muito o que fazer nessa área. Nesse sentido, extrapolar os resultados de experimentos computacionais para outras áreas do conhecimento, a fim de contribuir com a expansão de teorias gerais de consciência, devem ser feitos com muito cuidado. Esse é um caminho obscuro e que poderia levar facilmente um engenheiro a conclusões ingênuas com relação ao que é de fato consciência.

6.3 Trabalhos futuros

Com base no estudo e nos experimentos realizados, pode-se sugerir alguns trabalhos futuros:

- formalização do modelo de Baars-Franklin. Poderia-se utilizar uma rede de objetos (Gudwin, 1996) para isso;
- desenvolvimento de modelos de sensores reais de mercado, como modelos de simulação de câmeras digitais, infra-vermelho ou laser;
- melhorias no simulador, como adicionar a capacidade de simular diversos robos, e construir um ambiente de simulação 3D;
- adição de outros mecanismos presentes na ABF, como a inserção de memória associativa de longo prazo, memória episódica, mecanismos de metacognição e emoções;
- validação da ideia de criação de automatismos através do mecanismo de consciência, como sugere (Negatu, 2006). Para isso, parece importante o desenvolvimento das memórias de longo prazo e episódica;
- criação de um mecanismo de aprendizado de novos comportamentos para serem adicionados à rede de comportamentos durante a execução;
- desenvolvimento e teste de outros mecanismos de seleção de ação, como a rede de comportamentos estendida (Dorer, 1999);
- criação um editor e visualizador de redes de comportamento;
- desenvolvimento de uma arquitetura distribuída, para balancear a carga dos diversos codelets do agente, possibilitando a escalabilidade do sistema em um computador paralelo.

6.4 Considerações Finais

Duas abordagens paradigmáticas das ciências cognitivas são o *cognitivismo* e o *conexionismo*. No cognitivismo, predominante nas décadas de 70 e 80, a mente é vista como um computador serial, que manipula representações simbólicas e processa as informações sequencialmente (Wilson & Keil, 1999, p. 153-154). Contrapondo-se a essa teoria, com base na arquitetura paralela formada pelos neurônios do cérebro, o conexionismo prega que o fenômeno mental é descrito por meio de redes interconectadas de unidades simples, especializadas e muito eficientes. Essas unidades trabalham em paralelo, em uma rede, gerando um processo mental emergente (Medler, 1998). Não seria a consciência o mecanismo conciliador entre esses dois paradigmas aparentemente antagônicos?

Referências Bibliográficas

Amoroso, Richard L. 2003. The physical basis of qualia: overcoming the 1st person 3rd person barrier. *Noetic Journal*, **4**, 212–230.

Anwar, Ashraf, & Franklin, Stan. 2003. Sparse distributed memory for 'conscious' software agents. *Cognitive Systems Research*, **4**, 339–354.

Anwar, Ashraf, Dasgupta, Dipankar, & Franklin, Stan. 1999. Using Genetic Algorithms for Sparse Distributed Memory Initialization. *In: Proceedings of the 1999 Congress on Evolutionary Computation.*

Baars, Bernard J. 1988. *A cognitive theory of consciousness.* Cambridge University Press.

Baars, Bernard J. 1995. Can Physics Provide a Theory of Consciousness? A Review of Shadows of the Mind by Roger Penrose. *Psyche*, **2**(8), xx.

Baars, Bernard J. 1997. *In the Theater of Consciousness: The Workspace of the Mind.* Oxford University Press.

Baars, Bernard J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, **6**(1), 47–52.

Baars, Bernard J. 2003. Introduction: Treating Consciousness as a Variable: The Fading Taboo. *Chap. 1, pages 1–10 of:* Baars, Bernard J., Banks, William P., & Newman, James B. (eds), *Essencial Sources in the Scientific Study of Consciousness.* The MIT Press.

Baars, Bernard J., & Franklin, Stan. 2003. How conscious experience and working memory interact. *Trends in Cognitive Sciences*, **7**(4), 166–172.

Baars, Bernard J., & Franklin, Stan. 2007. An architectural model of conscious and unconscious brain functions: Global Workspace Theory and IDA. *Neural Networks*, **20**, 955–961.

Baars, Bernard J., Newman, James, & Taylor, J. G. 1998. Neuronal mechanisms of consciousness: A Relational Global Workspace framework. *Pages 269–278 of:* Hameroff, S., Kasznlak, A., & Laukes, J. (eds), *Toward a Science of Consciousness II: The second Tucson discussions and debates.* MIT Press.

Barsalou, Lawrence W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, **22**, 577–660.

Blackmore, Susan. 2003. Conciousness in Meme Machines. *Journal of Consciousness Studies*, **10**, 19–30.

Blackmore, Susan. 2005. *Consciousness - A very short introduction.* Oxford University Press.

Block, Ned. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, **2**, 227–287.

Block, Ned. 2002. Some Concepts of Consciousness. *In:* Chalmers, D. (ed), *Philosophy of Mind: Classical and Contemporary Readings.* Oxford University Press.

Blumberg, Bruce. 1994. Action-selection in Hamsterdam: Lessons from Ethology. *In: Proceedings of the Third International Conference on Simulation of Adaptive Behavior (SAB-94).*

Bogner, Myles, Maletic, Jonathan, & Franklin, Stan. 1999. *ConAg: A Reusable Framework for Developing Conscious Software Agents.*

Bogner, Myles Brandon. 1999 (December). *Realizing "Consciousness" in Software Agents.* Ph.D. thesis, The University of Memphis.

Brooks, Rodney A. 1986. A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, **RA-2**(1), 14–23.

Burgard, Wolfram, Cremers, Armin B., Fox, Dieter, Hähnel, Dirk, Lakemeyer, Gerhard, Schulz, Dirk, Steiner, Walter, & Thrun, Sebastian. 1999. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, **114**, 3–55.

Chalmers, David. 1995. Facing up the problem of consciousness. *Journal of Consciousness Studies*, **2**(3), 200–219.

Chalmers, David. 1996. *The conscious mind.* Oxford University Press.

Chalmers, David. 2002. What is a Neural Correlate of Consciousness? *In:* Noë, Alva, & Thompson, Evan (eds), *Visual and Mind - Selected Readings in the Philosophy of Perception.* MIT Press.

Chella, A., Liotta, M., & Macaluso, I. 2005 (October). CiceRobot, a cognitive robot for museum tours. *Pages 318–323 of: IASTED International Conference on Robotics and Applications.*

Chella, Antonio. 2007. Towards Robot Conscious Perception. *Pages 124–140 of: Artificial Consciousness.* Imprint Academic.

Chella, Antonio, & Macaluso, Irene. 2006. Sensations and Perceptions in "Cicerobot" a Museum Guide Robot. *In: Proceedings of BICS 2006.*

Conway, Martin A. 2001. Sensory-perceptual episodic memory and its context: autobiographical memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **356**(1413), 1375–1384.

Crick, Francis, & Koch, Christof. 1990. Towards a neurobiological theory of consciousness. *The Neurosciences*, **2**, 263–275.

Crick, Francis, & Koch, Christof. 1998. Consciousness and Neuroscience. *Cerebral Cortex*, **8**, 97–107.

Crick, Francis, & Koch, Christof. 2003. A framework for conciousness. *Nature Neuroscience*, **6**(2), 119–126.

Damasio, Antonio. 1998. Emotion in the perspective of an integrated nervous system. *Brain Research Reviews*, **26**, 83–86.

Damasio, Antonio. 1999. *The feeling of what happens - Body and Emotion in the Making of Consciousness.* Harverst Books.

Damasio, Antonio. 2000. *O mistério da consciência - Do corpo e das emoções ao conhecimento de si.* Companhia das Letras.

Dawkins, Richard. 1976. *The Selfish Gene.* Oxford University Press. 30th anniversary edition 2006.

de Almeida, Ana Maria Rocha, & El-Hani, Charbel Niño. 2006. Darwinismo Neural: Uma extensão metafórica da teoria da seleção natural. *Episteme*, **11**(24), 335–356.

de Morais Ribeiro, Henrique. 2001. Uma Revisão da Teoria Quântica da Consciência de Penrose e Hameroff. *Revista Eletrônica Informação e Cognição*, **3**(1), 108–125.

de Toro, Patrícia Rocha. 2007. *Sistemas de Controle Emocional Hedonista para Criaturas Artificiais.* M.Phil. thesis, Universidade Estadual de Campinas.

de Toro, Patrícia Rocha, Gudwin, Ricardo Ribeiro, & Miskulin, Mauro Sérgio. 2007 (Outubro). Agentes Emocionais Hedonistas para Comportamento Autônomo. *In: VIII SBAI - Simpósio Brasileiro de Automação Inteligente.*

Dehaene, Stanislas, & Changeux, Jean-Pierre. 2005. Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness. *Public Library of Science Biology*, **5**(10), e141.

Dehaene, Stanislas, Kerszberg, Michel, & Changeux, Jean-Pierre. 1998. A neuronal model of a global workspace in effortful cognitive tasks. *Pages 14529–14534 of: Proceedings of the National Academy of Sciences of the United States of America*, vol. 95.

Dehaene, Stanislas, Sergent, Claire, & Changeux, Jean-Pierre. 2003 (July). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Pages 8520–8525 of: Proceedings of the National Academy of Sciences of the United States of America*, vol. 100.

Dennett, Daniel C. 1969. *Content and Consciousness.* Routledge & Kegan Paul plc.

Dennett, Daniel C. 1991. *Conciousness Explained.* Back Bay Books.

Dennett, Daniel C. 1995. *Darwin's dangerous idea - Evolutions and the meanings of life.* Penguin Books.

Dennett, Daniel C. 2005. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness.* MIT Press.

Descartes, René. 1637. *Discurso do Método.* L & PM Pocket. Edição de 2005.

do Valle Filho, Adhemar Maria. 2003. *Um modelo para implementação de consciência em robôs móveis.* Ph.D. thesis, Universidade Federal de Santa Catarina.

Dorer, Klaus. 1999. Extended Behavior Networks for the magmaFreiburg Team. *Pages 79–83 of: In RoboCup-99 Team Descriptions for the Simulation League. Linkoping.*

Dubois, Daniel. 2007a (August). *Constructing an agent equipped with an artificial consciousness: application to an intelligent tutoring system.* Ph.D. thesis, Université du Québec à Montréal.

Dubois, Daniel. 2007b. What Does Consciousness Bring to CTS? *In: Consciousness and Artificial Intelligence: Theoretical foundations and current approaches - AAAI Symposium.* Association for the Advancement of Artificial Intelligence.

Dubois, Daniel, Nkambou, Roger, & Hohmeyer, Patrick. 2006. How "Consciousness" Allows a Cognitive Tutoring Agent Make Good Diagnosis During Astronauts' Training. *Lecture Notes in Computer Science*, **4053/2006**, 154–163.

e Silva, Maurício Marx, Álvares Fuhrmeister, Alida Vitória, Brum, Antônio Franscisco Maineri, Costa, Flávia, Rosito, Geraldo, Pizutti, Leandro Timm, Medeiros, Madeleine Scop, Ferreira, Paulo Picarelli, Breda, Renatu Lajús, Viero, Romulo, & Leite, Sérgio Silveira. 2003. A consciência: algumas concepções atuais sobre sua natureza, função e base neuroanatômica. *Revista de Psiquiatria do Rio Grande do Sul*, **25**(1), 52–64.

EDELMAN, GERALD M. 1978. Group selection and phasic re-entrant signalling: a theory of higher brain function. *In:* MOUNTCASTLE, V. B. (ed), *The mindful brain.* MIT Press.

EDELMAN, GERALD M. 1987. *Neural Darwinism: The theory of neuronal group selection.* Basic Books.

EDELMAN, GERALD M. 1992. *Bright Air, Brilliant Fire: On The Matter Of The Mind.* Basic Books.

EDELMAN, GERALD M. 2003. Naturalizing consciousness: A theoretical framework. *Proceedings of the National Academy of Sciences*, **100**(9), 5520–5524.

EDELMAN, GERALD M., & GALLY, JOSEPH A. 2001. Denegeracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, **98**(24), 13763–13768.

EDELMAN, GERALD M., & TONONI, GIULIO. 2000. *A Universe of Consciousness - How matter becomes imagination.* Basic Books.

FAYE, JAN. 2008. Copenhagen Interpretation of Quantum Mechanics. *In:* ZALTA, EDWARD N. (ed), *Stanford Encyclopedia of Philosophy.* Online.

FERNANDEZ-LEON, JOSE A., ACOSTA, GERARDO G., & MAYOSKY, MIGUEL A. 2009. Behavioral control through evolutionary neurocontrollers for autonomous mobile robot navigation. *Robotics and Autonomous System*, **57**, 411–419.

FRANKLIN, S., BAARS, B.J., RAMAMURTHY, U., & VENTURA, M. 2005. The Role of Consciousness in Memory. *Brains, Minds and Media*, **1**, 1–38.

FRANKLIN, STAN. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems*, **28**(6), 499–520.

FRANKLIN, STAN. 2003. IDA: A conscious artifact? *Journal of Consciousness Studies*, **10**(4-5), 47–66.

FRANKLIN, STAN, & FERKIN, MICHAEL. 2006. An Ontology for Comparative Cognition: A Functional Approach. *Comparative Cognition & Behavior Reviews*, **1**, 36–52.

FRANKLIN, STAN, & GRAESSER, ART. 1999. A software agent model of consciousness. *Consciousness and Cognition*, **8**(September), 285–301.

FRANKLIN, STAN, & JR, F. G. PATTERSON. 2006. The LIDA architecture: adding new modes of learning to an intelligent, autnomomous, software agent. *In: Integrated Design and Process Technology, IDPT-2006.*

FRANKLIN, STAN, KELEMEN, ARPAD, & MCCAULEY, LEE. 1998 (October). IDA: a cognitive agent architecture. *Pages 2646–2651 of: IEEE Conference on Systems.* Systems, Man, and Cybernetics, vol. 3.

Franklin, Stanley. 2005. A "Conciousness" Based Architecture for a Functioning Mind. *Chap. 8, pages 149–175 of:* Davis, Darryl N. (ed), *Visions of Mind: Architecture for Cognition and Affect.* Idea Group Inc (IGI).

Franklin, Stanley P. 1995. *Artificial Minds.* The MIT Press.

Friendlander, David, & Franklin, Stan. 2008. LIDA and a theory of mind. *In:* Goertzel, Ben, & Wang, Pei (eds), *Proceedings of AGI-08.* IOS Press.

Gaglio, Salvatore. 2007. Intelligent Artificial Systems. *Pages 97–115 of:* Chella, Antonio, & Manzotti, Riccardo (eds), *Artificial Consciousness.* Imprint Academic.

Gamez, David. 2008a. *The Development and Analysis of Conscious Machines.* Ph.D. thesis, University of Essex.

Gamez, David. 2008b. Progress in Machine Consciousness. *Consciousness and Cognition*, **17**, 887–910.

Goldberg, Robert P. 1974. Survey of Virtual Machine Research. *IEEE Computer*, **7**(6), 34–45.

González, Fernando M. Montes, Hernández, Antonio Marín, & Figueroa, Homero Ríos. 2006. An Effective Robotic Model of Action Selection. *Lecture Notes in Computer Science*, **4177**, 123–132. Selected Papers from the 11th Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2005).

Gudwin, Ricardo Ribeiro. 1996. *Contribuições ao Estudo Matemático de Sistemas Inteligentes.* Ph.D. thesis, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas.

Güzeldere, Güven. 1997. The Many Faces of Consciousness: A Field Guide. *In:* Block, Ned, Flanagan, Owen, & Güzeldere, Güven (eds), *The Nature of Consciousness: Philosophical Debates.* MIT Press.

Hagan, S., Hameroff, Stuart, & Tusynski, Jack. 2002. Quantum computation in brain microtubules: Decoherence and biological feasibility. *Physical Review*, **65**, 61901.

Haikonen, Pentti. 2000a. An Artificial Mind via Cognitive Modular Neural Architecture. *In: Proceedings Symposium on How to Design a Functioning Mind AISB00 Convention.*

Haikonen, Pentti 0 A. 2000b (July). A modular neural system for machine cognition. *Pages 47–50 of: Proceedings of the IEEE-INNS-ENNS International Joint Conference*, vol. 1.

Haikonen, Pentti O. 2003. *The cognitive approach to conscious machines.* Imprint Academic.

HALFPAP, DULCE MARIA. 2005. *Um modelo de consciência para aplicação em artefatos inteligentes.* Ph.D. thesis, Universidade Federal de Santa Catarina.

HAMEROFF, STUART. 1987. *Ultimate Computing.* Elsevier.

HAMEROFF, STUART. 1998. Quantum computation in brain microtubules? The Penrose-Hameroff 'Orch OR' model of consciousness. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **356**, 1869–1896.

HAMEROFF, STUART. 2007. The Brain Is Both Neurocomputer and Quantum Computer. *Cognitive Science*, **31**, 1035–1045.

HAMEROFF, STUART, & PENROSE, ROGER. 1996. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation*, **40**, 453–480.

HAMEROFF, STUART, & PENROSE, ROGER. 2003. Conscious events as orchestrated space-time selections. *NeuroQuantology*, **1**, 10–35.

HAMEROFF, STUART, NIP, ALEX, PORTER, MITCHELL, & TUSYNSKI, JACK. 2002. Condution pathways in microtubules, biological quantum computation, and consciousness. *BioSystems*, **64**, 149–168.

HOFSTADTER, DOUGLAS R.; MELANIE MITCHELL. 1994. The Copycat Project: A model of mental fluidity and analogy-making. *In Holyoak, K.J & Barnden, J.A. (Eds.). Advances in connectionist and neural computation theory*, **2**, 31–112.

HOLLAND, OWEN, & KNIGHT, ROB. 2006. The Anthropomimetic Principle. *In:* BURN, JEREMY, & WILSON, MYRA (eds), *Proceedings of the AISB06 symposium on biologically inspired robotics.*

HOLLAND, OWN, KNIGHT, ROB, & NEWCOMBE, RICHARD. 2007. A Robot-based Approach to Machine Consciousness. *Pages 156–173 of:* CHELLA, ANTONIO, & MANZOTTI, RICARDO (eds), *Artificial Consciousness.* Imprint Academic.

HUI, NIRMAL BARAN, & PRATIHAR, DILIP KUMAR. 2009. A comparative study on some navigation shemes of a real robot tackling moving obstacles. *Robotics and Computer-Integrated Manufacturing*, **25**, 810–828.

INABA, KEITA, & TAKENO, JUNICHI. 2003. Consistency between Cognition and Behavior Creates Consciousness. *Systemics, Cybernetics and Informatics*, **2**(4), 52–57.

JACKSON, JOHN V. 1987. Idea for a mind. *ACM SIGART Bulletin*, **xx**(101), 23–26.

JAMES, WILLIAM. 1904. Does "Consciousness" Exist? *Journal of Philosophy, Psychology, and Scientific Methods*, **1**, 477–491.

KANERVA, PENTTI. 1988. *Sparse Distributed Memory.* MIT Press.

KANERVA, PENTTI. 1991 (March). *Efficient Packing of Patterns in Sparse Distributed Memory by Selective Weighting of Input Bits.* Tech. rept. 9108. Research Institute for Advanced Computer Science NASA Ames Research Center.

KIRSH, DAVID. 1990. When is information explicitly represented? *Pages 341–365 of:* HANSON, P. (ed), *Information, language, and cognition.* Vancouver, Canada: University of British Columbia Press.

LEWIS, CLARENCE IRVING. 1929. *Mind and the World Order: Outline of a Theory of Knowledge.* New York: C. Scribner's Sons.

LIOTTA, M., CHELLA, A., INGRAFFIA, N., MACALUSO, I., PILATO, G., & VASSALLO, G. 2005 (September). A Semantic Information Retrieval in a Robot Museum Guide Application. *In: Proceedings of AI*IA Workshop for Cultural Heritage.*

LLINÁS, R., U., RIBARY, CONTRERAS, D., & PEDROARENA, C. 1998. The neuronal basis for consciousness. *Philosophical Transactions: Biological Sciences*, **353**(1377), 1841–1849.

LLINÁS, RODOLFO. 2003. Consciousness and the thalamocortical loop. *International Congress Series*, **1250**, 409–416. Cognition and emotion in the brain. Selected topics of the International Symposium on Limbic and Association Cortical Systems.

LOW, KIAN HSIANG, LEOW, WEE KHENG, & JR, MARCELO H. ANG. 2004. Continuous-Spaced Action Selection for Single- and Multi-Robot Tasks Using Cooperative Extended Kohonen Maps. *Pages 198–203 of: Proceedings of the 2004 IEEE lnternational Conference on Netwarkinp, Sensing & Control Taipei. Taiwan, March 21-23.*

MAES, PATTIE. 1989. How to do the right thing. *Connection Science Journal*, **1**, 3.

MCCAULEY, THOMAS LEE, & FRANKLIN, STAN. 2002. A Large-Scale Multi-Agent System for Navy Personnel Distribution. *Connection Science*, **14**, 371–385.

MCFETRIDGE, L., & IBRAHIM, M. Y. 2009. A new methodology of mobile robot navigation: The agoraphilic algorithm. *Robotics and Computer-Integrated Manufacturing*, **25**, 545–551.

MEDLER, DAVID A. 1998. A Brief History of Connectionism. *Neural Computing Surveys*, **1**(2), 18–72.

MILLER, GEORGE A. 1962. *Psychology: The Science of Mental Life.* Penguin Books.

MINSKY, MARVIN. 1986. *The society of mind.* New York, NY: Simon & Schuster.

MOURA, IVAN. 2006. A model of agent consciousness and its implementation. *Neurocomputing*, **69**, 1984–1995.

NAGEL, THOMAS. 1974. What is it like to be a bat? *Pages 159–174 of:* LYONS, WILLIAM (ed), *Modern Philosophy of Mind.* Everyman.

Negatu, Aregahegn Seifu. 2006 (August). *Cognitively Inspired Decision Making for Software Agents: Integrated Mechanisms for Action Selection, Expectation, Automatization and Non-Routine Problem Solving.* Ph.D. thesis, The University of Memphis.

Negatu, Aregahegn Seifu, & Franklin, S. 2002. An Action Selection Mechanism for C̈onsciousS̈oftware Agents. *Cognitive Science Quarterly*, **2**, 363–386.

Nii, H. Penny. 1986. The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures. *AI Magazine*, **7**(2), 38–53.

Nuxoll, Adrew M. 2007. *Enhancing Intelligent Agents with Episodic Memory.* Ph.D. thesis, University of Michigan.

Penrose, Roger. 1989. *The emperor's new mind.* Oxford University Press.

Penrose, Roger. 1994. *Shadows of Mind.* Oxford University Press.

Pinto, Hugo Silva Corrêa. 2005. *Designing Autonomous Agents for Computer Games with Extended Behavior Networks: An Investigation of Agent Performance, Character Modeling and Action Selection in Ureal Tournament.* M.Phil. thesis, Universidade Federal do Rio Grande do Sul.

Ramamurthy, U., D'Mello, S. K., & Franklin, Stan. 2004. Modified sparse distributed memory as transient episodic memory for cognitive software agents. *Pages 5858– 5863 of: IEEE International Conference on Systems, Man and Cybernetics*, vol. 6.

Ramamurthy, Uma, Bogner, Myles, & Franklin, Stan. 1998. Conscious Learning In An Adaptive Software Agent. *Pages 24–27 of: Proceedings of The Second Asia Pacific Conference on Simulated Evolution and Learning.*

Ramamurthy, Uma, Baars, Bernard J., D'Mello, Sidney K., & Franklin, Stan. 2006. LIDA: A Working Model of Cognition. *Pages 244–249 of:* Danilo Fum, Fabio Del Missier, & Stocco, Andrea (eds), *Proceedings of the 7th International Conference on Cognitive Modeling.* Edizioni Goliardiche.

Robertson, Edwin M. 2007. The serial reaction time task: implicit motor skill learning? *The Journal of Neuroscience*, **27**(38), 10073–10075.

Rogers, David. 1988. *Kanerva's Sparse Distributed Memory: An Associative Memory Algorithm Well-Suited to the Connection Machine.* Tech. rept. Research Institute for Advanced Computer Science - NASA Ames Research Center.

Ros, Raquel, Arcos, Josep Lluís, de Mantaras, Ramon Lopez, & Veloso, Manuela. 2009. A case-based approach for coordinated action selection in robot soccer. *Artificial Intelligence*, **173**, 1014–1039.

Sallas, Bill, Mathews, Robert C., Lane, Sean M., & Sun, Ron. 2007. Developing rich and quickly accessed knowledge of an artificial grammar. *Memory and Cognition*, **35**(8), 2118–2133.

Sanz, Ricardo, López, Ignacio, & Bermejo-Alonso, Julita. 2007. A Rationale and Vision for Machine Consciousness in Complex Controllers. *Pages 141–155 of:* Chella, Antonio, & Manzotti, Riccardo (eds), *Artificial Consciousness*. Imprint Academic.

Seager, William. 2007. A Brief History of the Philosophical Problem of Consciousness. *In:* Zelazo, Philip David, Moscovitch, Morris, & Thompson, Evan (eds), *The Cambridge Handbook of Consciousness*. Cambridge University Press.

Searle, John. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, **3**(3), 417–457.

Searle, John. 1997. *The Mystery of Consciousness*. New York Review of Books.

Selfridge, Oliver G. 1958 (November). Pandemonium: a paradigm for learning. *Pages 513–526 of: Mechanism of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory.*

Seth, Anil K., & Baars, Bernard J. 2005. Neural Darwinism an consciousness. *Consciousness and Cognition*, **14**, 140–168.

Shadlen, Michael N., & Movshon, J. Antony. 1999. Synchrony Unbound: A Critical Evaluation of the Temporal Binding Hypothesis. *Neuron*, **24**, 67–77.

Shanahan, Murray. 2006. A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition*, **15**, 433–449.

Shanahan, Murray. 2007. A spiking neuron model of cortical broadcast and competition. *Consciousness and Cognition*, **17**, 288–303.

Shanahan, Murray, & Connor, Dustin. 2008. Modeling the Neural Basis of Cognitive Integration an Consciousness. *In: Artificial Live XI, Conference on the Simulation and Synthesis of Living Systems.*

Song, Hongjun. 1998 (May). *Control Structures for Software Agents*. Ph.D. thesis, The University of Memphis.

Song, Hongjun, & Franklin, Stan. 2000. A behaviour instantiation agent architecture. *Connection Science*, **12**(1), 21–44.

Sugerman, Jeremy, Venkitachalam, Ganesh, & Lim, Beng-Hong. 2001. Virtualizing I/O Devices on VMware Workstation's Hosted Virtual Machine Monitor. *In: Proceedings of the 2001 USENIX Annual Technical Conference.*

Sun, Ron. 1995. Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, **75**, 241–295.

Sun, Ron. 1997. Learning, Action and Consciousness: A Hybrid Approach Toward Modelling Consciousness. *Neural Networks*, **10**(7), 1317–1331.

Sun, Ron. 1999. Accounting for the Computational Basis of Consciousness: A Connectionist Approach. *Consciousness and Cognition*, **8**, 529–565.

Sun, Ron. 2003. *A tutorial on CLARION*. Tech. rept. Rensselaer Polytechnic Institute. http://www.cogsci.rpi.edu/ rsun/sun.tutorial.pdf. Visitado em 04.11.2008.

Sun, Ron. 2007. The Challenges of Building Computational Cognitive Architectures. *Pages 37–60 of: Challenges for Computational Intelligence*. Studies in Computational Intelligence, vol. 63/2007. Springer Berlin / Heidelberg.

Sun, Ron, & Franklin, Stan. 2007. Computational Model of Consciousness: A Taxonomy and Some Examples. *Pages 151–174 of:* Zelazo, Philip David, Moscovitch, Morris, & Thompson, Even (eds), *The Cambridge Handbook of Consciousness*. Cambridge University Press.

Sun, Ron, & Naveh, Isaac. 2004. Simulating organizational decision-making using a congnitively realistic agent model. *Journal of Artificial Societies and Social Simulation*, **7**(3), online. http://jasss.soc.surrey.ac.uk/7/3/5.html Acessado em: 10.11.2008.

Sun, Ron, & Naveh, Isaac. 2007. Social Institution, Cognition, and Survival: A Cognitive-Social Simulation. *Mind and Society*, **6**(2), 15–142.

Sun, Ron, & Peterson, Todd. 1998. Some experiments with a hybrid model for learning sequential decision making. *Information Sciences*, **111**, 83–107.

Sun, Ron, & Peterson, Todd. 1999. Multi-agent reinforcement learning: weighting and partitioning. *Neural Networks*, **12**, 727–753.

Sun, Ron, & Sessions, Chad. 2000. Self-Segmentation of Sequences: Automatic Formation of Hierarchies of Sequential Behaviors. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics*, **30**(3), 403 – 418.

Sun, Ron, & Terry, Chris. 2002. Implicit Learning of Serial Reaction Time Tasks: Connectionist vs. Symbolic Models. *In: Proceedings of the 24th Annual Conference of the Cognitive Science Society.*

Sun, Ron, Peterson, Todd, & Merrill, Edward. 1996. Bottom-up skill learning in reactive sequential decision tasks. *Pages 684–690 of: Proceedings of 18th Cognitive Science Society Conference.*

Sun, Ron, Merrill, Edward, & Peterson, Todd. 2001. From implicit skills to explicit knowledge: a bottom-up model of skill learning. *Cognitive Science*, **25**, 203–244.

Sun, Ron, Terry, Chris, & Slusarz, Paul. 2005. The interaction of the explicit and implicit in skill learning: a dual process approach. *Psychological Review*, **112**(1), 159–192.

Sun, Ron, Zhang, Xi, & Mathews, Robert. 2009. Capturing human data in a letter counting task: Accessibility and action-centeredness in representing cognitive skills. *Neural Networks*, **22**, 15–29. in press.

Suárez, Lizet Liñero. 2000. *Conhecimento Sensorial - Uma Análise segundo a perspectiva da Semiótica Computacional.* M.Phil. thesis, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas.

Suzuki, Tohru, Inaba, Keita, & Takeno, Junichi. 2005. Conscious Robot That Distinguishes Between Self and Others and Implements Imitation Behavior. *In: Innovations in Applied Artificial Intelligence - 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems.* Lecture Notes in Computer Science. Bari, Italy: Springer Berlin / Heidelberg.

Takeno, Junichi, Inaba, Keita, & Suzuki, Tohru. 2005 (June). Experiments and examination of mirror image cognition using a small robot. *In: Proceedings 2005 IEEE International Symposium on Computational Intelligence in Robotics and Automation.*

Tononi, Giulio, & Edelman, Gerald M. 1998. Consciousness and Complexity. *Science*, **282**, 1846–1851.

Tulving, Endel. 2002. Episodic Memory: From Mind to Brain. *Annual Review of Psychology*, **53**(February), 1–25.

Tye, Michael. 2007 (July). *Qualia.* Stanford Encylopedia of Philosophy. Online em http://plato.stanford.edu/entries/qualia/. Acessado em 27.10.2008.

Tyrrell, Toby. 1993. *Computational Mechanisms for Action Selection.* Ph.D. thesis, University of Edinburgh.

Vannini, Antonella. 2008. Quantum Models of Consciousness. *Quantum Biosystems*, **2**, 165–184.

Watkins, Christopher John Cornish Hellaby. 1989. *Learning from delayed rewards.* Ph.D. thesis, King's College.

Wilson, Robert A., & Keil, Frank (eds). 1999. *The MIT Encyclopedia of the Cognitive Sciences.* MIT Press.

Woolf, Nancy J., & Hameroff, Stuart. 2001. A quantum approach to visual consciousness. *Trends in Cognitive Sciences*, **5**(11), 472–478.

Young, Robert M. 1990. The Mind-Body Problem. *Pages 702–711 of:* C., Olby R. (ed), *Companion to the History of Modern Science.* Routledge.

Zeman, Adam. 2008. Consciousness: concepts, neurobiology, terminology of impairments, theoretical models and philosophical background. *Disorders of Consciousness*, **90**, 3–31.

Zhang, Zhaohua, Franklin, Stan, Olde, Brent, Graesser, Art, & Wan, Yun. 1998. Natural language sensing for autonomous agent. *In: Proceedings of IEEE International Joint Symposia on Intelligence and Systems.*