UNIVERSIDADE FEDERAL DO CEARÁ
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

CARLOS ROBERTO RODRIGUES FILHO

A EVOLUÇÃO DO PROJETO COMPUTACIONAL PARA UMA INTELIGÊNCIA ARTIFICIAL E AS NOVAS PERSPECTIVAS OFERECIDAS PELOS AVANÇOS DA COGNIÇÃO ENATIVA

FORTALEZA, CEARÁ

2012

CARLOS ROBERTO RODRIGUES FILHO

A EVOLUÇÃO DO PROJETO COMPUTACIONAL PARA UMA INTELIGÊNCIA ARTIFICIAL E AS NOVAS PERSPECTIVAS OFERECIDAS PELOS AVANÇOS DA COGNIÇÃO ENATIVA

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal do Ceará, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Carlos Eduardo Fisch de Brito

FORTALEZA, CEARÁ

2012

A000z Rodrigues Filho, C. R..

A Evolução do Projeto Computacional para uma Inteligência Artificial e as Novas Perspectivas Oferecidas pelos Avanços da Cognição Enativa / Carlos Roberto Rodrigues Filho. 2012.

67p.;il. color. enc.

Orientador: Prof. Dr. Carlos Eduardo Fisch de Brito

Co-Orientador:

Dissertação(Ciência da Computação) - Universidade Federal do Ceará, Departamento de Computação, Fortaleza, 2012.

1. 2. 3. I. Prof. Dr. Carlos Eduardo Fisch de Brito(Orient.) II. Universidade Federal do Ceará– Ciência da Computação(Mestrado) III. Mestre

CDD:000.0

RESUMO

O desenvolvimento da área de Inteligência Artificial (IA) trouxe grandes avanços para a resolução de problemas computacionalmente difíceis. No entanto, a meta inicial da IA, de implementação de uma inteligência genuína de forma artificial, ainda não foi alcançada. Por isso, a grande maioria dos pesquisadores da área mudou o foco de suas pesquisas para resolução de problemas, em geral abandonando a meta de uma inteligência artificial. Para continuar perseguindo essa meta, outros pesquisadores passaram a questionar os princípios filosóficos da IA e tomar novos rumos. Alguns deles perceberam que o processo da inteligência não é puramente mental. De acordo com essa nova visão, o corpo tem um papel fundamental no processo cognitivo. A partir desse ponto de vista, visando superar obstáculos tradicionais da IA, surgiu a IA Corporificada. Esta tem uma forte tendência a implementação em robôs, para se desenvolver um melhor projeto sobre o corpo.

Porém, apesar de ter obtido avanços em relação a problema da IA tradicional, a IA Corporificada começou a apresentar suas próprias limitações. Surgiu então a ideia de que um agente genuinamente inteligente deve formular seus próprios problemas a partir da percepção da realidade, construída em termo do seus aparato sensório-motor. Em outras palavras, a inteligência genuína está ligada à autonomia do agente. A IA Enativa surgiu influenciada por estudos biológicos a respeito da autonomia. Nessa área da IA a meta é construir um agente artificial autônomo.

Este trabalho relata a trajetória da IA desde a sua fundação, passando pela vertente da IA Corporificada, e apontando um possível novo paradigma da IA Enativa. Além disso, nós Analisamos e discutimos os processos que levaram pesquisadores a questionar o embasamento filosófico da IA e a formular novos conceitos a respeito do que é inteligência.

Palavras-chave: Inteligência Artificial, Enação, IA Corporificada, Autonomia, Autopoiese.

ABSTRACT

The development in the Artificial Intelligence (AI) field brought great improvement to the resolution of computationally hard problems. However, the early goal of AI, of implementing a genuine intelligence in an artificial way, was not achieved. Therefore most of the field's researchers changed their research's focus to problem resolution, quitting the goal of an artificial intelligence. To keep pursuing that goal others researchers started questioning the philosophical principles of AI and they took new routes. Some of them realized that the process of intelligence is not purely mental. According to this new view the body has a fundamental role in the cognitive process. From this point of view, aiming to overcome traditional obstacles of AI, the Embodied AI emerged. It has a strong tendency to implementation in robots, to develop a better design of the body.

Despite having achieved improvements over the problem of traditional AI, Embodied AI started to present its own limitations. Then came the idea that a genuinely intelligent agent must formulate its own problems from the perception of reality, constructed in terms of their sensory-motor apparatus. In other words, the genuine intelligence is linked to the agent's autonomy. The Enactive AI appeared influenced by biological studies about autonomy. In this AI field the goal is to build an autonomous artificial agent.

This paper reports the trajectory of AI since its foundation, through strand of Embodied AI, and pointing to a possible new paradigm of Enactive AI. In addition, we analyse and discuss the processes that led researchers to question the philosophical basis of IA and formulate new concepts about what intelligence is.

Keywords: Artificial Intelligence, Enaction, Embodied AI, Autonomy, Autopoiesis.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O fenômeno da cognição é o objeto de estudo de muitas áreas do conhecimento como a Ciência Cognitiva, a Filosofia, a Psicologia e a Inteligência Artificial. A Ciência Cognitiva e parte da Filosofia têm a inteligência como objeto de estudo em si mesma, no sentido de que elas buscam encontrar uma explicação para o advento da cognição. Por outro lado, a Psicologia e a Inteligência Artificial buscam encontrar no entendimento da cognição a possibilidade de entender o que é essencial ao fenômeno, para a Psicologia, ou entender o funcionamento do mesmo, para a Inteligência Artificial. Essa diversidade de pontos de vista sobre o mesmo tema levou a elaboração de diferentes concepções sobre o que seja cognição ou inteligência, e não se pode dizer hoje em dia que exista um definição rigorosa e consensual para a noção de inteligência. Esse fato é claramente ilustrado por Legg e Hutter(LEGG; HUTTER, 2007), que reuniram um grande número de definições de inteligência formuladas por pesquisadores de diversas áreas. Um fato que chama a atenção é que mesmo quando nos restringimos ao escopo da inteligência artificial, não encontramos consenso. Pesquisadores diferentes colocam ênfase em aspectos diferentes do fenômeno da inteligência. Por exemplo, Minsky identifica como característica essencial de um agente inteligente "*... the ability to solve hard problems*."[1](MINSKY, 1988) *apud* (LEGG; HUTTER, 2007). Já Goertzel destaca o ambiente em que o agente está inserido ao dizer que o comportamento inteligente consiste em "*Achieving complex goals in complex environments*."[2](GOERTZEL, 2006) *apud* (LEGG; HUTTER, 2007).

As diferenças entre as diversas concepções de inteligência dentro da área de Inteligência Artificial (IA) se evidenciam na atual coexistência de pelo menos três vertentes distintas de pesquisa. A primeira vertente, que chamaremos de IA Tradicional ou IA Simbólica e Representacionalista, está associada aos primeiros desenvolvimentos da área e é a vertente dominante até hoje. Ela se caracteriza por afirmar que a inteligência é uma manifestação exclusiva da mente, baseada na manipulação de representações mentais da realidade. A segunda vertente é chamada de IA Corporificada[3] (*Embodied Artificial Intelligence*), e surgiu em contraponto à IA Representacionalista. Seu ponto de vista é baseado em uma série de estudos que evidenciam o papel fundamental do corpo no processo cognitivo(BATESON, 1979). A área tem um caráter essencialmente experimental, tendo encontrado na robótica os elementos necessários para a instanciação do corpo dos agentes artificiais. A terceira vertente, chamada IA Enativa(FROESE; ZIEMKE, 2009), ainda se encontra em seus primeiros desenvolvimentos e tem poucos trabalhos práticos, mas se distingue da anterior por uma forte caracterização conceitual. A IA Enativa representa um avanço com relação à IA Corporificada, ao exigir que o agente inteligente seja capaz de gerar os seus objetivos a partir de uma perspectiva própria sobre o mundo. A chave para essa possibilidade estaria na ideia de que a identidade do agente precisa ser autonomamente constituída.

É importante observar que a perspectiva enativa introduz um novo elemento con-

[1] ... a habilidade de resolver problemas difíceis.

[2] Alcançar metas complexas em ambientes complexos.

[3] O termo *embodied*, aplicado à inteligência artificial, não possui uma tradução para o português amplamente estabelecida. Neste trabalho, optamos por usar o termo inteligência artificial corporificada.

ceitual nas discussões sobre cognição e inteligência artificial. Ela oferece o diagnóstico de que a falta de autonomia dos seres artificiais compromete a meta de se conseguir uma inteligência artificial genuína. A autonomia não é um conceito simples, tendo significados específicos em áreas diferentes. Por exemplo, no contexto dos dispositivos que utilizam bateria, ela se refere ao tempo de atividade do dispositivo sem a necessidade de se conectar à rede elétrica. No contexto da automação, o conceito de autonomia pode ser entendido como a capacidade que um mecanismo possui de realizar uma tarefa sem o auxílio de um ser humano. A noção de autonomia da IA Enativa é um pouco mais sutil, e diz respeito à capacidade que um agente artificial deve ter de gerar seus próprios objetivos. Essa ideia parece paradoxal à primeira vista: como é que uma máquina (computacional) cujo comportamento é estritamente especificado em termo de regras pré-definidas, pode gerar suas próprias regras e objetivos? Essa questão nunca foi colocada seriamente nas discussões da IA. Apesar da ideia de agentes inteligentes autônomos estar sempre presente no imaginário da área, os projetos concretos da IA Tradicional sempre envolveram a construção de artefatos que alcançam objetivos definidos pelos seus construtores.

1.1 O Conceito de inteligência na IA Tradicional

Tradicionalmente, o conceito de inteligência para a IA sempre esteve ligado à ideia de um agente ser capaz de resolver um problema. Além disso, a alta capacidade humana de resolução de problemas foi adotada como meta, de modo que um agente artificial deve ser capaz de resolver problemas com uma habilidade comparável à de um ser humano para que seja considerado inteligente. Por exemplo, um programa de xadrez que não consegue vencer sequer uma criança não é considerado inteligente. No entanto, essa impressão se modifica se o programa pode vencer um bom jogador.

Não é difícil programar um computador para resolver problemas matemáticos ou problemas relacionados a jogos de tabuleiro. O que esses problemas têm em comum é o fato de possuírem uma definição formal e um escopo bem delimitado. No entanto, em princípio, a IA se propõe a resolver problemas de todo tipo, baseada na observação de que a inteligência é uma faculdade genérica. A dificuldade em viabilizar esse projeto está no fato de que a maior parte dos problemas resolvidos por um ser humano cotidianamente não possui uma especificação explícita. Por exemplo, uma pessoa que acorda e deseja ficar pronto para sair para ir ao trabalho, note que não existe uma maneira única de se fazer isso, nem tampouco uma só solução. Para que esses problemas possam ser resolvidos por um computador, eles precisam passar por um processo de formalização.

O primeiro passo no processo de formalização consiste em abstrair do problema todos os seus aspectos irrelevantes do ponto de vista computacional. O resultado desse passo é a identificação de um conjunto de propriedades que permitem descrever o problema em termos de um objetivo a ser alcançado pela execução de ações bem definidas que modificam o estado do mundo. Uma vez formalizado, o problema pode ser representado na forma de uma estrutura de dados adequada para a manipulação por um computador, e a sua solução pode ser encontrada pela execução de um algoritmo. Continuando o exemplo acima, imagine alguém que deseje ir de casa para o trabalho, o primeiro passo para o computador resolver este problema seria

colocar uma restrição do tipo "utilizando o caminho mais curto", para se definir qual das várias soluções deve ser buscada. Em seguida é preciso modelar as ruas através de um grafo, deve-se também estabelecer as regras para "caminhar"nesse grafo e os pontos inicial e final.

Uma observação importante a ser feita nesse momento é que, na IA Tradicional, o processo de formalização não é considerado como uma parte da resolução do problema, no sentido de que ele apenas captura o que é essencial em uma dada situação, e transforma um problema real em um problema tratável pelo computador. Mais precisamente, a IA Tradicional se baseia no pressuposto de que a inteligência pode ser reduzida à resolução de um 'núcleo' lógico do problema. McCarthy ilustra bem esse ponto, ao definir inteligência da seguinte maneira: "*Intelligence is the computational part of the ability to achieve goals in the world*."[4](MCCARTHY, 2007) *apud* (LEGG; HUTTER, 2007).

Esse esquema de formalização de problemas e codificação da sua solução na forma de um algoritmo foi responsável por um grande número de sucessos nos primeiros anos da IA[5]. No entanto, quando examinado mais de perto, esse esquema não parece oferecer uma estratégia satisfatória para o projeto da IA, pois o computador não participa da definição do problema e nem tampouco da elaboração do algoritmo. Um mecanismo que se limita a realizar operações combinatórias predefinidas pelo programador sobre estruturas de dados desprovidas de significado, dificilmente poderia ser considerado como algo inteligente.

Alguns avanços posteriores da IA reduzem um pouco o impacto dessa crítica, ao permitirem que certos aspectos da definição do problema, e da sua solução, não sejam previamente especificados pelo programador. Por exemplo, a área de meta-heurísticas investiga métodos para a solução de problemas de otimização através de procedimentos iterativos, com base em uma função de avaliação. O aspecto relevante aqui é que esses métodos não utilizam nenhuma informação estrutural do problema que está sendo resolvido, e podem ser aplicados a uma larga classe de problemas. Outro exemplo importante é fornecido pela área de aprendizado automático. Nas técnicas de aprendizado por reforço o computador não possui uma descrição explícita do objetivo que deve ser alcançado, e deve aprimorar a qualidade da solução que ele tem no momento através de um processo de tentativa e erro utilizando as ações que ele tem à sua disposição. Essas ideias estão presentes na seguinte definição de inteligência formulada por Nakashima:

> Intelligence is the ability to process information properly in a complex environment. The criteria of properness are not predefined and hence not available beforehand. They are acquired as a result of the information processing.[6](NAKASHIMA, 1999) *apud* (LEGG; HUTTER, 2007)

Como o ser humano é a referência de inteligência utilizada pela IA Tradicional, uma outra maneira de avaliar o sucesso do projeto de implementar uma inteligência artificial seria

[4]Inteligência é a parte computacional da habilidade de atingir metas no mundo.

[5]Como mostraremos no Capítulo 2

[6]Inteligência é a habilidade de processar informação corretamente em um ambiente complexo. Os critérios de correção não são pré-definidos e portanto não estão disponíveis antecipadamente. Eles são adquiridos como um resultado do processamento de informação.

comparar os modos pelos quais os seres humanos e o computador resolvem problemas. O ser humano é plenamente capaz de lidar com a novidade, sendo capaz de resolver um problema que nunca foi encontrado antes com base na sua experiência passada. Já o computador só consegue lidar com a novidade se ela tiver sido antecipada pelo programador. Por exemplo, considere o jogo de xadrez e um computador programado para jogá-lo. Tanto o ser humano quanto o computador em questão são capazes de jogar xadrez. Agora imagine que se quer jogar uma variante do xadrez onde o peão só se move no sentido diagonal. O ser humano é plenamente capaz de aprender a jogar esse novo jogo. O computador, por sua vez, deve ser reprogramado para poder jogar. Mas, podemos imaginar um programa em que as regras sobre a movimentação das peças sejam um dos parâmetros do programa. Esse programa conseguiria jogar as variantes do xadrez que utilizam o mesmo conjunto peças. No entanto, se for introduzida uma peça nova, este computador também terá que ser reprogramado. Se pensarmos em sucessivas modificações no jogo, podemos perceber que não há limite para as possibilidades do computador, contanto que ele sempre seja previamente programado. Qualquer modificação não antecipada requer uma reprogramação. Ou seja, em princípio, não há limite para a variedade de comportamentos que um programa de computador pode apresentar. Mas, uma vez que o programa esteja definido, ele não pode adquirir um novo comportamento. Essa característica de determinação prévia, que é intrínseca aos programas de computador, é chamada de anterioridade. Essa propriedade é problemática, pois ela estabelece um limite fundamental para a capacidade de resolução de problemas de qualquer programa. Se todas as possibilidades de ação do programa precisam ser especificadas *a priori*, então sempre haverá um limite para elas, e em face de um fato imprevisto o problema não será resolvido. Esse limite parece não existir para os seres humanos, que aparentam estar sempre aptos para lidar com o novo e o desconhecido.

O problema da anterioridade surge como resultado do processo de formalização. Ao identificar os aspectos do problema que são relevantes para a sua solução, a formalização acaba deixando de lado tudo o que não está diretamente ligado à resolução do problema. Ou seja, em um certo sentido, todas as possibilidades de ação para a resolução do problema já estão presentes no ato de formalização. Podemos notar que a formalização envolve um processo de escolha entre as várias soluções possíveis para um problema. O que fica claro a partir dessas observações é que a tarefa de formalização de um problema é também, ela própria, um problema que precisa ser resolvido. No esquema da IA Tradicional, essa tarefa é sempre realizada por um ser humano. Isso significa que o computador implementa apenas uma etapa do processo de resolução de problemas. De fato, a etapa menos nobre: a manipulação mecânica de símbolos. Nesse sentido, os programas da IA Tradicional devem ser considerados como meros instrumentos acessórios dos seres humanos na sua atividade inteligente de resolução de problemas.

É importante observar que essa situação não se modifica de maneira essencial com os desenvolvimentos posteriores da IA mencionados acima. Podemos enxergar nas técnicas de meta-heurística, aprendizado automático e outras, uma tentativa de eliminar, ou pelo menos diminuir, o problema da anterioridade. No entanto, tudo o que se consegue com esses movimentos é apenas mudar a dificuldade de lugar. Em todos esses casos, temos algoritmos especificados em um nível de abstração mais alto, onde certos aspectos dos problemas que são resolvidos não aparecem explicitamente, mas que utilizam critérios de qualidade ou sucesso especificados pelo programador ou usuário do programa.

A questão que permanece então é se é possível remover a anterioridade de fato. A única solução possível parece consistir em fazer com que o próprio computador identifique o que é relevante em uma dada situação. Alguns pesquisadores sugeriram que o corpo deve ter um papel crucial nessa tarefa, uma vez que ele é o elemento mediador entre o agente e a realidade. São as características físicas e sensoriais do corpo que determinam as possibilidades de ação do agente, e seria através de um processo de experimentação e avaliação que ele descobriria os meios para chegar a solução do problema. Essas observações levaram a investigação para o campo da robótica, onde é possível realizar experimentos com sistemas computacionais concretos (robôs) que interagem com o ambiente através de sensores e motores.

1.2 Mudanças de Paradigma na IA

Se nos distanciarmos um pouco do movimento feito pela IA, podemos ter uma visão mais ampla dos caminhos que a área seguiu e enxergar mudanças mais profundas ao longo do seu desenvolvimento. Tais mudanças se deram quando parte dos pesquisadores da área se voltaram para outro paradigma de inteligência, iniciando uma nova vertente da IA. Apesar de apresentarmos agora essas mudanças de forma breve, elas foram fruto do próprio desenvolvimento de cada vertente que as antecedem. A discussão detalhada desse desenvolvimento é o que constitui o corpo deste trabalho.

No início, a IA era baseada no paradigma de que inteligência é igual a resolução de problemas formalizados. Como vimos, a formalização é um processo que reduz o problema a seu núcleo lógico, retirando o problema do seu contexto. No fim do processo de formalização, o que resta são símbolos e regras para a manipulação desses símbolos, representado pelas ações e seus pré-requisitos. A IA Tradicional, ou IA Simbólica, deu origem à IA Corporificada. O paradigma da IA Corporificada é que o corpo tem um papel importante no processo cognitivo. A ideia de que inteligência é resolução de problemas ainda é bastante presente na IA Corporificada. Porém, traçando um paralelo com a IA simbólica, os problemas da IA Corporificada não são formalizados, no sentido de que o agente da IA Corporificada deve ser capaz de captar no seu ambiente a "configuração" do mundo em que se encontra, também de identificar se as pré-condições para a execução de suas ações são satisfeitas. O objetivo, todavia, ainda é idealizado pelo seu construtor. Já para a IA Enativa, o agente genuinamente inteligente deve ser capaz de gerar seus próprios objetivos. Apesar de caracterizarmos a IA Enativa como vertente distinta da IA Corporificada, o pouco tempo desde que ela surgiu não nos permite afirmar se irá se confirmar essa distinção. Atualmente todas as três subáreas da IA estão ativamente sendo pesquisadas e discutidas entre seus pesquisadores. Podemos dizer que o que distingue as três é sua interpretação do que é inteligência e como ela se manifesta.

O estabelecimento da IA simbólica se deu a partir da década de 1950, quando Minsky e McCarthy organizaram a conferência de Dartmouth de 1956, ao perceber que outros pesquisadores também estavam utilizando o computador na tentativa de automatizar as capacidades cognitivas. Apesar do esforço de organização da área, somente mais tarde, na década de 1970, surgiu uma teoria abrangente da IA Simbólica, quando Newell e Simon propuseram uma hipótese forte a respeito da cognição. A Hipótese do Sistema Simbólico Físico, sobre a qual

discutiremos na Seção 2.2, diz que um sistema simbólico físico possui os meios necessários e suficientes para a ação inteligente(NEWELL; SIMON, 1976). Segundo essa hipótese, nós dispomos de um sistema simbólico interno que é responsável pela nossa capacidade cognitiva. Na ciência cognitiva, essa hipótese serviu de base para o estabelecimento do chamado paradigma computacional. A aceitação da hipótese de Newell e Simon fora da computação teve impacto na IA, reforçando a ideia de que a área seguia o rumo correto, mesmo tendo recebido críticas, das quais falaremos a seguir.

A demora para o surgimento de uma teoria da IA Simbólica foi efeito de uma área que sempre se caracterizou por ser pragmática. A ideia geral era que se era possível reproduzir o fenômeno da inteligência no computador, então o estabelecimento de teorias pode ficar em segundo plano. Porém, importantes limitações se apresentaram aos avanços da área. Como a dificuldade de resolver versões maiores de problemas cujas versões menores eram facilmente resolvidas, como o problema da torre de hanoi, por exemplo. O crescimento exponencial da dificuldade de resolução dos problemas, em relação ao tamanho da entrada, foi uma das limitações que se apresentou como empecilho para a IA. Esse tipo de limitação ganhou um reforço teórico com o surgimento da Teoria da Complexidade. Outro tipo de problema que se apresentou como difícil para a IA foram os relacionados com linguagem natural. Alguns filósofos, como Searle, afirmaram a impossibilidade de entendimento de linguagem natural por parte de um computador, como discutiremos em detalhes na Seção 2.4.1. Todas essas limitações e argumentos a respeito de impossibilidades para a IA fizeram com que alguns pesquisadores da IA se voltassem para discussões teóricas sobre o embasamento da área.

No início da IA, alguns filósofos acusavam os pesquisadores da área de estarem refazendo o que a filosofia já havia feito, isto é, de não aproveitarem a discussão filosófica já existente a respeito da inteligência. Porém, a discussão de inteligência na filosofia, apesar de estar relacionada com a IA, está em outro nível, no sentido de que aborda aspectos diferentes da inteligência, e não tem, em princípio, implicação direta na IA. Por exemplo, Dreyfus faz sua crítica à IA (DREYFUS, 1975) com base na filosofia de Heidegger, mas o texto original do autor alemão, não trata especificamente da noção de inteligência. Portanto, era natural que os pesquisadores da IA não enxergassem na discussão filosófica, que foram acusados de ignorar, uma relação com os seus projetos. Depois de certo tempo, as críticas filosóficas aumentaram e alguns dos próprios cientistas da IA passaram a considerá-las, a fim de compreender melhor as barreiras que impediam o desenvolvimento da IA. Nesse ponto, o conceito de representação, que é uma noção fundamental para a IA Simbólica, foi apontado como o grande empecilho para o desenvolvimento de uma inteligência genuína.

Das discussões a respeito da representação surge uma nova vertente de IA, chamada de IA Corporificada. A IA Corporificada inicia como uma oposição ao aspectos representacionalista da IA Simbólica. O corpo passa a ser entendido como parte do processo cognitivo, daí o nome IA Corporificada. A inteligência não é mais entendida como um processo que acontece isoladamente no cérebro, mas que acontece de forma paralela no corpo todo. Dessa forma, a meta de inteligência da IA Corporificada também é diferente, no sentido de que, em princípio, não se busca a inteligência humana como meta, mas a inteligência de um animal. Este passa a ser considerado também como referência de inteligência.

Com uma ideia de cognição descentralizada, essa área da IA aproxima a percepção da ação, através da interação direta entre sensores e motores. Dessa forma o agente entra em contato direto com o mundo. Como Brooks afirma “It turns out to be better to use the world as its own model.”[7](BROOKS, 1991), ao invés de um modelo mental. Brooks toma como referência a teoria da evolução e diz que a natureza demorou muito mais tempo para evoluir até seres da complexidade de insetos, do que desse nível até seres humanos. Brooks pretende, então, através de um projeto incremental de inteligência, perseguir a meta da inteligência no nível de um inseto, para daí chegar ao nível dos seres humanos.

As implementações de IA Corporificada são preferencialmente feitas com robôs, ou simulação de robôs. Por estarem embutidos no mundo, os robôs superam algumas dificuldades da IA Tradicional. Por exemplo, na IA Simbólica o desenvolvedor deveria definir o nível de detalhamento e os aspectos relevantes do mundo para resolver o problema, através da formalização. Como na IA Corporificada não há representação, todos os níveis de detalhes, em princípio, estão disponíveis para o agente, limitados pela capacidade de seus sensores. Mesmo em uma simulação implementada de acordo com a IA Corporificada, os robôs são implementados com sensores, que são a única forma de eles adquirirem informações sobre seu ambiente. Nesse caso, precisamos ser cautelosos para não confundirmos a representação que a simulação em si é, com uma possível representação que o agente simulado teria do seu mundo. Uma simulação em um computador é, sem dúvidas, uma representação daquilo que é simulado. Porém, um agente corporificado simulado não necessariamente é programado para ter acesso a todas as informações do seu ambiente, ele somente percebe os aspectos do ambiente com os quais seus sensores entraram em contato. Por exemplo, considere uma simulação de um agente que deve aprender a cruzar um labirinto. Todas as informações a respeito do labirinto fazem parte da simulação, mas o agente não tem acesso a elas, ou seja, ele não pode consultar o mapa do labirinto e fazer cálculos em cima dele. O agente só tem acesso ao que está diretamente em contato com seus sensores, e terá que utilizar seus motores para modificar as posições dos seu sensores, a fim de conseguir mais informações a respeito do seu ambiente.

A IA Corporificada surgiu como uma tentativa de superar as barreiras da IA Simbólica. Note que o problema da anterioridade, na forma que estava presente na IA Simbólica, não existe na IA Corporificada. Isto é, o problema que surge ao se tentar antecipar todas as possibilidades que o agente venha a se deparar, não está presente em um agente que tem acesso direto ao mundo, sem utilizar representações. Ele deve ser capaz de aprender a lidar com as variações do problema a medida que elas surgem. Podemos perceber a anterioridade como um limitante às possibilidades do agente da IA. Ao tentar retirar anterioridade, tentamos dar mais "liberdade" ao agente. Note que estabelecer um problema para o agente resolver é, de certa forma, limitá-lo. Ou seja, estabelecer sua finalidade, e construí-lo com um propósito pré-determinado, ainda caracteriza a presença da anterioridade. Todo o projeto do robô, desde aspectos físicos até aspectos lógicos são pensados com o intuito de resolver um problema. Ao fazer isso, estamos ainda antecipando o comportamento do agente.

A solução para o problema da anterioridade na IA Corporificada seria produzir um agente completamente autônomo, capaz de gerar suas próprias metas. Com esse intuito surge

[7] Acontece que é melhor usar o mundo como seu próprio modelo.

a vertente da IA Enativa. Antes de o conceito de enação chegar a IA, ele já existia no contexto da Ciência Cognitiva. Varela, Thompson e Rosch (VARELA; THOMPSON; ROSCH, 1992) puseram a pedra fundamental desse ramo da Ciência Cognitiva. Por se tratar de uma área recente, ainda não é possível afirmar com clareza se a IA Enativa constitui uma vertente por si só, que estabelecerá um novo paradigma, ou se é uma subárea da IA Corporificada. Nesse trabalho nós tratamos como áreas distintas, pelo fato de a IA Enativa possuir uma característica bem distinta da IA Corporificada, conceitualmente falando.

1.3 Enação

O paradigma da enação está fundamentado sobre a ideia de que mesmo o nível mais simples de cognição só pode se manifestar em um agente autônomo. A Teoria da Enação se sustenta sobre a teoria biológica da autopoiese. De acordo com essa teoria, um ser vivo é caracterizado por estar em constante processo de autoprodução. A autoprodução é o procedimento através do qual as partes do ser vivo produzem umas às outras em uma rede interdependente. Ao produzirem uma as outras, as partes delimitam o que faz parte do organismo e o que não faz. Assim, a autoprodução constante do ser vivo garante uma identidade ao mesmo. Note que o processo de autoprodução é executado pelo ser vivo e tem como produto o próprio ser vivo. Podemos dizer que esse processo possui em si mesmo o propósito intrínseco de continuidade da autoprodução. Outro modo de falar sobre esse propósito é dizer que o ser vivo tem o propósito de continuar vivo. A geração intrínseca de um propósito caracteriza o ser vivo como um agente que segue suas próprias leis, ou seja, que é autônomo. A partir do propósito de manutenção da autoprodução e da diferenciação entre dentro e fora do organismo vivo, o ser vivo gera uma perspectiva que lhe permite avaliar suas interações com seu meio. Isto é, uma interação com o meio pode ser julgada como boa ou ruim para a auto-manutenção. Essa capacidade de diferenciar as interações a partir de um critério, podemos chamar de cognição no seu modo mais simples.

Um aspecto importante do ponto de vista da enação é a percepção de que o mundo se apresenta para cada indivíduo de forma diferente, dependendo do seu aparato sensório-motor. Dessa forma, indivíduos com aparatos sensoriais parecidos, como dois representantes de uma mesma espécie, percebem o mundo de forma similar. Por outro lado, um indivíduo com visão percebe o mundo de forma diferente de outro que não tenha essa capacidade. Por exemplo, dois seres humanos, em geral, possuem o mesmo espectro de percepção de cores, já o cão possui um espectro mais limitado, por possuir um aparato sensorial diferente. O aparato motor do indivíduo determina a forma como ele pode agir no seu meio. Da mesma forma que acontece com o aparato sensorial, indivíduos com aparatos motores semelhantes podem agir no mundo de maneiras semelhantes. A separação dos aparatos sensorial e motor só é feita aqui para fins de explicação, na realidade os dois influenciam tanto na percepção como na ação. Esse entendimento do papel do aparato sensório-motor na determinação do laço percepção-ação já se encontrava na IA Corporificada.

Voltando à ideia do ser autônomo capaz de julgar suas interações com o meio, o modo como ele interage com o meio determina a relevância de aspectos externos a ele na rea-

lização de seus objetivos. Assim, podemos perceber que apesar de o meio poder ser o mesmo, para indivíduos autônomos diferentes, ele se apresenta de formas diferentes. Por isso, dizemos que o ser autônomo constrói sua própria percepção de mundo. A percepção de mundo construída pelo agente autônomo é subjetiva, no sentido de que cada aspecto depende do contexto em que o agente se encontrava no momento de sua interação. Assim, a percepção construída pelo agente autônomo não é uma representação da realidade objetiva.

Na IA, o paradigma da Enação coloca a questão da construção da percepção como central na síntese de um agente artificial inteligente. Como essa percepção não pode ser construída de fora do agente, pelo seu projetista por exemplo, o agente deve ser projetado de modo a ter condições de realizar tal construção da percepção. Portanto, a IA Enativa se volta para a questão de como produzir um agente autônomo, capaz de ter uma percepção de mundo própria e, desse modo, gerar também seus próprios objetivos. Para tanto nos voltamos para a discussão a respeito de autonomia em seres vivos, feita por Varela em (VARELA, 1991). Mas primeiro, precisamos discutir teoria da autopoiese, de Maturana e Varela, que caracteriza o ser vivo como um ser que se autoproduz. Essa discussão será aprofundada no Capítulo 4

1.4 Organização do Trabalho

O presente trabalho se propõe a apresentar e a discutir as modificações sofridas pela IA no seu rumo a área da Inteligência Artificial Enativa. Para tanto, iniciaremos discutindo a área da Inteligência Artificial, como foi concebida, no Capítulo 2, onde apresentaremos os fundamentos filosóficos e matemáticos que ajudaram a formar a área na Seção 2.1. Em seguida, na Seção 2.2, mostraremos o surgimento da área propriamente dita e seus primeiros desenvolvimentos. Posteriormente, na Seção 2.3, apresentaremos a mudança ocorrida na IA tradicional e como ela se encontra nos dias de hoje. Finalizando o capítulo, discutiremos as críticas feitas à área na Seção 2.4.

No Capítulo 3, apresentaremos a vertente de IA denominada IA Corporificada. Inicialmente, faremos uma discussão a respeito de como a IA Corporificada surge para contrapor pontos da IA tradicional que dificultam a produção de inteligência genuína em seres artificiais baseados em computador. Na Seção 3.1, mostraremos como Brooks iniciou a área com suas ideias de fazer IA sem representação. Já na Seção 3.2, abordamos como as ideias de corporificação da inteligência influenciou a Ciência Cognitiva, e como isso influenciou de volta a IA. A IA Corporificada será aprofundada na Seção 3.3, com alguns de seus trabalhos discutidos na Seção 3.4. Finalmente, apresentaremos as críticas à IA Corporificada na Seção 3.5.

No Capítulo 4, discutiremos as propostas que surgiram para superar as críticas à IA Corporificada e como isso nos leva ao estudo de teorias da biologia. Na Seção 4.1 apresentaremos a teoria da autopoiese. Mostraremos algumas de suas implicações filosóficas, discutidas pelos próprios autores da teoria, na Seção 4.2. Na Seção 4.3 veremos a visão de Di Paolo a respeito das consequências da autopoiese, apresentando mais uma característica dos seres vivos que ele considera essencial para cognição: adaptatividade. Finalmente, na Seção 4.4 discutiremos a respeito de todos esses conceitos que envolvem o paradigma da Enação no contexto da

Inteligência Artificial. No Capítulo 5 faremos as conclusões do trabalho e falaremos a respeito dos trabalhos futuros.

2 DESENVOLVIMENTO E CRÍTICA DA INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial, tradicionalmente, se fundamenta no pressuposto de que é possível compreender o mundo dividindo-o em partes e representando-as através de um sistema de símbolos. Segundo essa visão, nós pensamos através de uma representação do mundo. Ou seja, nós teríamos, dentro da nossa mente, representações dos fatos do mundo que manipulamos e sobre as quais raciocinamos. O raciocínio é apenas a manipulação de regras sobre as representações que fazemos do mundo. Apreender essas regras tem sido o objetivo dos cientistas da IA, que pretendem em seguida implementá-las em uma máquina. As representações do mundo, juntamente com as regras do raciocínio, fariam da máquina uma entidade inteligente. Ela estaria preparada para agir nas situações em que já dispusesse das devidas representações.

Nesse capítulo, apresentaremos as bases de pensamento que geraram os pressupostos assumidos pela IA, na Seção 2.1. Na Seção 2.2, falaremos do surgimento da IA como área independente, seus desenvolvimentos, suas dificuldades e a mudança de metas de alguns cientistas da IA. Na Seção 2.3, mostraremos os desdobramentos dessa mudança nas metas da IA. Finalmente, na Seção 2.4, apresentaremos as principais críticas feitas ao projeto da Inteligência Artificial.

2.1 Herança filosófica da Inteligência Artificial

Russell e Norvig(RUSSELL; NORVIG, 2010) fazem um resumo da história da IA começando por Aristóteles, que em seu *Organon* publicou as primeiras ideias sobre a lógica, ainda de maneira informal, e que foram a base para a atual área de lógica. Ele também estava procurando identificar as regras básicas para o raciocínio. Muito tempo depois veio Thomas Hobbes, que em 1651 no livro Leviatã sugeriu a possibilidade de um animal artificial, mecânico. Hobbes acreditava que todo raciocínio é apenas computação numérica. Nessa época, Blaise Pascal já havia construído a famosa Pascaline (Figura 2.1), uma máquina de calcular capaz de realizar as operações de adição e subtração. Até então muitos acreditavam que o cálculo era uma faculdade exclusiva do pensamento humano, e que, portanto, somente seres inteligentes eram capazes de calcular. Isso fez com que Pascal imaginasse que sua máquina estava mais

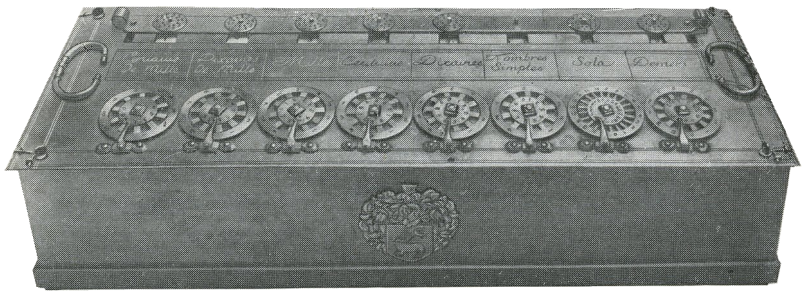

Figura 2.1: Máquina da Calcular Produzida por Pascal.
Fonte: http://en.wikipedia.org/wiki/File:Pascaline_calculator_front.png

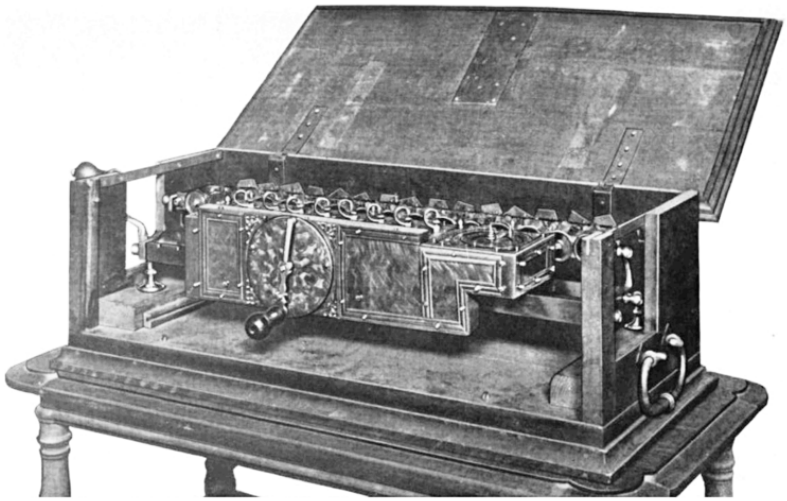

Figura 2.2: Máquina da Calcular Produzida por Leibniz.
Fonte: http://history-computer.com/MechanicalCalculators/Pioneers/Lebniz.html

próxima do pensamento do que um animal. Depois, Gottfried Leibniz construiu outra calculadora mecânica (Figura 2.2), esta capaz também de multiplicar, dividir e obter raízes. Diversos aspectos do raciocínio estavam sendo reproduzidos em máquinas, o que já levava a crer que seria possível fazer o mesmo com a inteligência.

Russell e Norvig(RUSSELL; NORVIG, 2010) continuam, explicando que René Descartes fez uma das primeiras discussões filosóficas sobre mente e matéria, distinguindo as duas. Descartes desenvolve sua filosofia ortogonalmente à tendência das máquinas de calcular. Descartes afirma que as coisas (*res*) mentais têm uma natureza distinta das coisas físicas. Para Descartes as coisas físicas estão fadadas ao determinismo, por estarem sujeitas às leis da Física. Mas, para ele, colocar a mente na mesma condição tiraria o seu livre arbítrio. Por isso Descartes separa a coisa física (*res extensa*) da coisa mental ou pensante (*res cogitans*). Assim, o sujeito é constituído de duas partes de natureza distinta: uma pensante e outra física. Essa teoria, de que a nossa essência está na mente e que o nosso corpo é somente uma ferramenta que usamos para interagir com o mundo, se tornou a visão dominante e natural na sociedade ocidental. Assim, ela também serviu de base para a IA formular seus pressupostos de que os aspectos físicos do mundo são representados na parte pensante do ser.

Finalmente, no século 20, surgiu a teoria da computação, um modelo teórico de máquina programável capaz de computar funções que não estavam predeterminadas, funções arbitrárias. A possibilidade de programar uma máquina deu mais força às ideias de Hobbes, que já estavam imersas na cultura popular através da ficção científica (BUCHANAN, 2005). A capacidade matemática da máquina programável remetia à ideia de que a mesma era inteligente, ou seja, tinha uma capacidade mental compatível com o ser humano, e assim poderia executar tarefas compatíveis com o intelecto humano. Só seria necessário, então, representar o mundo dentro da máquina, como se acreditava acontecer com os seres humanos. Essa comparação com a inteligência humana fez crescer a crença de que os seres humanos pensam como os computadores, ou seja, através de cálculos, o que mais tarde fez surgir a hipótese de que um

sistema simbólico é suficiente para se obter inteligência, da qual falaremos na Seção 2.2.

2.2 Início da Inteligência Artificial

A conferência de Dartmouth em 1956, organizada por Marvin Minsky e John McCarthy, é tida como o marco que simboliza o início da área de Inteligência Artificial. De fato, aí que o nome da área foi cunhado por McCarthy(CREVIER, 1993)(MCCARTHY et al., acessado em 2012). Os dois cientistas tiveram a ideia de reunir pesquisadores que estivessem estudando o desenvolvimento de tarefas inteligentes pelo computador. Estavam presentes cientistas que se tornaram proeminentes pesquisadores da área nas suas primeiras décadas, como Ray Solomonoff, Oliver Selfridge, Trenchard More, Arthur Samuel, Allen Newell e Herbert Simon. Conforme podemos observar em (MCCARTHY et al., acessado em 2012), o otimismo era grande e perceptível em afirmações como "*The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it*."[1] A proposta era que fossem discutidos na conferência temas como a simulação de funções cerebrais, o uso de linguagem pelo computador, redes neurais, e a teoria da complexidade de funções. Esse último tema tinha como foco lidar com funções exponenciais. Ainda, entre os temas, estavam o autoaperfeiçoamento das máquinas, a classificação de abstrações, aleatoriedade e criatividade.

Na época da conferência, Newell e Simon já trabalhavam em um programa de prova automática de teoremas, o *Logic Theorist*, que pouco depois já era capaz de provar a maioria dos teoremas do segundo capítulo do *Principia Mathematica* de Bertrand Russell e Alfred Whitehead. O programa foi implementado na primeira linguagem de programação desenvolvida para a IA, a IPL (*Information Processing Language*) que utilizava os mesmos princípios de Lisp. Mais tarde, Newell e Simon desenvolveram um outro programa, que seguiria protocolos humanos para a resolução de problemas, o *General Problem Solver* (GPS). O GPS talvez tenha sido o primeiro programa explicitamente construído com o intuito de pensar como os seres humanos. O GPS deveria resolver qualquer problema formalizado simbolicamente, tendo conseguido resolver problemas como a torre de Hanoi, mas esbarrando na explosão combinatória dos problemas.

A figura 2.3 mostra um diagrama de estados do GPS. Seu funcionamento em alto nível é relativamente simples. Ele avalia a distância em que se encontra do seu objetivo. Se ele ainda não tiver atingido o seu objetivo, ele encontra um meio de reduzir a distância até o seu objetivo. Então, verifica se ele tem as condições necessárias para utilizar esse meio. Se não tiver, resolve esse problema com uma chamada recursiva do GPS. Uma vez que ele tem as condições para utilizar o meio, o faz e retorna para o ponto inicial, onde verifica se chegou ao seu objetivo.

O sucesso do GPS fez com que os autores, em 1976, formulassem a Hipótese do Sistema Simbólico Físico (*Physical Symbol System Hypothesis*), que diz " *a physical symbol*

[1] O estudo se dará com base na conjectura de que todo aspecto do aprendizado, ou qualquer outra característica da inteligência, pode, em princípio, ser tão precisamente descrito que uma máquina pode ser construída para simulá-lo.

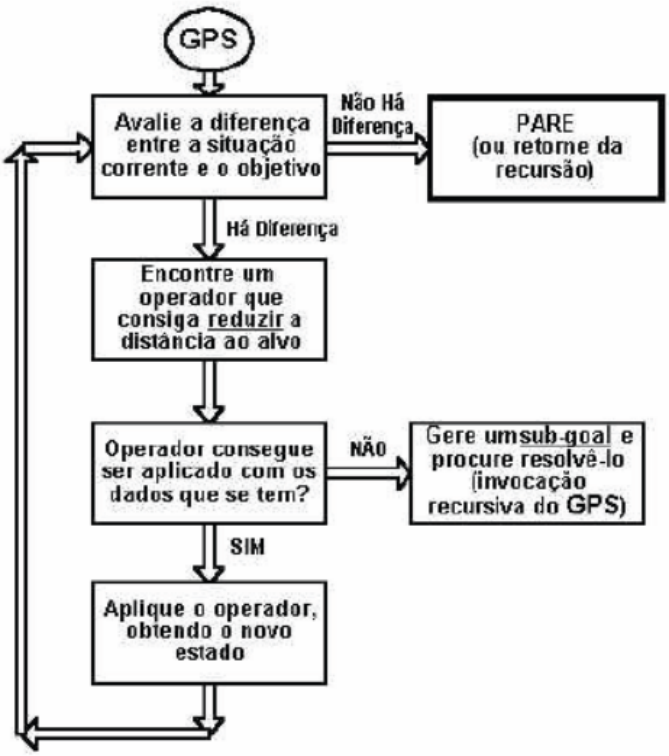

Figura 2.3: Diagrama de estados do GPS

system has the necessary and sufficient means for general intelligent action" [2](RUSSELL; NORVIG, 2010)(NEWELL; SIMON, 1976). Se essa hipótese for verdadeira, isso significa que o cérebro tem a mesma estrutura lógica do computador e que ele também seria um sistema simbólico físico. Apesar de muitas críticas, como mostraremos na Seção 2.4, essa hipótese se tornou o fundamento conceitual para o desenvolvimento da IA da época. A hipótese também fez com que essa vertente da IA ficasse conhecida como IA Simbólica.

2.2.1 Outros desenvolvimentos

Desde 1952 Arthur Samuel já vinha trabalhando em um programa para o jogo de damas, que conseguia jogar melhor que o seu criador usando heurísticas. O programa chegou a ser apresentado em um programa de televisão em 1956, causando uma forte impressão na população americana. Em 1957 H. Simon afirmou que dentro de dez anos um computador seria capaz de vencer um grande mestre de xadrez[3]. Não é difícil ver que os jogos de tabuleiro constituem uma boa classe de problemas para a IA simbólica, já que, normalmente, eles têm regras bem definidas e possuem um escopo limitado. Uma estratégia simples para um jogador automático consiste em verificar todas as possíveis implicações futuras de um determinado movimento. Isso pode ser feito através da aplicação das regras do jogo. Por exemplo, no xadrez, o computador calcula os possíveis movimentos do adversário se ele próprio, computador, fizer determinado movimento. Se em todos os cenários, ou a maioria, o computador sofrer uma

[2]Um sistema simbólico físico dispõe dos meios necessários e suficientes para ação inteligente.
[3]Levou cerca de 40 anos para que ocorresse.

derrota, ele descarta aquela jogada como possível. Porém, as possibilidades crescem exponencialmente, o que limita a previsão do computador. Os jogos eletrônicos modernos normalmente utilizam técnicas de IA para interagir com os jogadores.

Em 1958, John McCarthy publicou sua linguagem de programação Lisp, que dominou a IA nos 30 anos subsequentes. Na época, a linguagem dominante era o Fortran. McCarthy inovou com o Lisp ao introduzir a estrutura condicional *if-then-else*, em Fortran só havia condicionais com *goto*. O uso de funções em Fortran era obscuro, muito mais parecido com um condicional. Em Lisp as funções tinham uma representação literal, ficando separadas do código principal e tornando o código mais claro. Além disso, Lisp foi a primeira linguagem a implementar a recursão. Em resumo, o Lisp modificou radicalmente a maneira de se programar computadores e por isso se tornou tão popular e ainda hoje é utilizada, inclusive dentro de programas que não tem relação direta com programação, como programas de desenho assistido por computador, por exemplo.

Ainda em 1958, McCarthy lançou um artigo no qual descrevia um programa hipotético, chamado *Advice Taker*, que seria um sistema completo de inteligência artificial. O programa usaria o cálculo de predicados para derivar conclusões a partir de uma lista de premissas. Dessa vez ele não deveria chegar somente às conclusões matemáticas ou da própria lógica, como o *Logic Theorist*, mas acerca de qualquer assunto.

Enquanto isso, outros cientistas trabalhavam para que a inteligência computacional avançasse em outras direções. Um exemplo é o trabalho de Terry Winograd *Blocks World*, que consiste em um pequeno mundo virtual formado de blocos sobre uma mesa, que podem ser movidos através de um vocabulário limitado de comandos em linguagem natural. O objetivo original de Winograd era o estudo da linguagem natural. Mas, o seu programa permitiu a outros pesquisadores elaborarem estudos a respeito de visão e aprendizado. Outros exemplos de trabalhos em outras direções são os trabalhos sobre raciocínio, como o trabalho de Slagle(SLAGLE, 1963) para resolver integrais e o de Evans(EVANS, 1964) que reconhecia padrões em testes de quociente de inteligência.

2.2.2 Desenvolvimentos posteriores

No final da década de 1960, surgiu um novo tipo de programa na IA. Um sistema especialista se propunha a reunir o conhecimento completo de um especialista em uma área pontual. Tentava-se, pela primeira vez, organizar as regras não formais utilizadas por um ser humanos na realização de uma tarefa específica, para permitir à máquina imitar o comportamento humano de tomada de decisão. O mais famoso exemplo de sistema especialista é o MyCin(YU et al., 1979), desenvolvido na linguagem de programação de McCarthy, o Lisp. O sistema MyCin era capaz de diagnosticar infecções por bactéria e receitar tratamentos com antibióticos. Apesar do MyCin ter tido uma média de acertos acima da obtida pelos médicos que se submeteram aos mesmos testes (YU et al., 1979), ele nunca foi usado na prática por razões éticas.

Na década de 1960, também, ganhou força a vertente de IA que ficou conhecida

como *conexionista*. A IA conexionista se baseia na ideia de que um modelo suficientemente completo e preciso do cérebro humano apresentará comportamento inteligente. Para a construção de tal modelo do cérebro, os cientistas utilizam modelos matemáticos de neurônios. A partir do modelo matemático, é construído um neurônio artificial, computacionalmente ou através de circuitos elétricos. Estes devem ser interconectados, formando uma rede neural, daí o nome conexionista. Depois de formada a rede, esta deve ser treinada, para que as conexões entre os neurônios se ajustem e ela seja capaz de resolver o problema para o qual foi concebida.

A partir de 1966 grandes dificuldades começaram a se apresentar para a área da IA, resultando em cortes de verbas governamentais(RUSSELL; NORVIG, 2010). Algumas dessas dificuldades estavam relacionadas com o fracasso da tentativa de tradução automática do russo para o inglês, e com os problemas encontrados para ampliar o escopo dos programas que antes eram limitados. Com o otimismo abalado, muitos pesquisadores começaram a mudar o foco de seus trabalhos. As pesquisas de IA se mostravam promissoras para escopos restritos, e então alguns passaram a não almejar muito mais do que aquilo que seus programas já faziam. A IA então ficaria focada em resolver problemas específicos. O desenvolvimento da teoria da complexidade também foi um fator importante nessa mudança, já que estabeleceu limites teóricos que antes não eram vistos pelos cientistas.

2.3 A Inteligência Artificial moderna

A partir da mudança de rumo da IA mencionada na seção anterior, a IA se apresentou à indústria como a solução para alguns problemas de otimização de processos industriais e de distribuição. Também passou a ser aplicada para produzir pilotos automáticos em veículos e na robótica. Nas aplicações fora da pesquisa, a IA sempre esteve desvinculada da ideia de produção de inteligência de modo artificial. Isso também se refletiu na pesquisa da IA simbólica e conexionista, de modo geral.

Hoje, a IA está inserida em nosso cotidiano. No início da década de 2000, muitos acreditavam não ser possível criar uma ferramenta de busca eficiente para a internet (DREYFUS, 2009), por causa da não organização das páginas e do crescimento rápido da rede. No entanto, Winograd, que já havia sido convencido pelas críticas que veremos na Seção 2.4, orientou um aluno de doutorado que conseguiu desenvolver uma ferramenta de busca eficiente chamada Google.

Atualmente nos cursos de IA são apresentadas técnicas de busca, que podem ser aplicadas em otimização. Algoritmos genéticos, que utilizam a ideia da evolução, o processo para chegar em uma solução ótima, ou próximo do ótimo, é divididos em etapas chamadas gerações. Cada geração apresenta um conjunto de soluções para o problema, seleciona-se as melhores a partir de um critério pré-determinado, probabilisticamente ou não. As soluções selecionadas são combinadas para formar as soluções da geração seguinte. Além desses, outro tópico comum nos cursos de IA são as redes neurais, que comumente são combinadas com os algoritmos genéticos.

Na pesquisa, podemos destacar algumas áreas de atuação da IA. A resolução de

problemas, aplicada a jogos e sistemas especialistas. O raciocínio por senso comum, que utiliza mecanismos de inferência e base de dados de conhecimento de senso comum. Tais bases de dados são preenchidas por pessoas comuns. Reconhecimentos de visão e fala, como mecanismos de reconhecimento biométrico. Processamento de linguagem natural, que produz regras para a automatização do entendimento da linguagem natural, podendo servir também para tradução automática.

Como podemos ver, a IA se modificou bastante. Em grande parte, deixou de lado propósito de produção de máquinas inteligentes e passou a se concentrar na resolução de problemas que não são triviais para os computadores. Com isso, ampliou bastante sua atuação e passou a servir como uma ferramenta essencial para o desenvolvimento tecnológico.

2.4 Crítica da Inteligência Artificial

Apesar do avanço da IA no desenvolvimento de técnicas para a resolução de problemas difíceis para a computação, seu projeto inicial, de implementar a inteligência de forma artificial, nunca chegou a ter êxito pleno, somente pontual. Os primórdios da IA foram caracterizados por um pragmatismo muito forte, seus sucessos pareciam não necessitar de discussões conceituais, pelo menos na visão dos próprios pesquisadores de IA. Com o surgimento das dificuldades que mostramos na Seção 2.2, os próprios cientistas da IA começaram a se questionar sobre que teorias a respeito da inteligência embasavam suas pesquisas. Eles se voltaram para críticas que já existiam. Estas feitas por filósofos, que viam na IA uma área promissora para se testar teses filosóficas que até então eram só teóricas (DREYFUS, 1975). Porém, os pressupostos que embasavam a pesquisa de IA foram duramente criticados.

A filosofia da Inteligência Artificial não se ocupa em analisar sucessos ou fracassos específicos, mas se interessa em analisar quais os pressupostos filosóficos/conceituais que estão por trás dos projetos da IA. Toda teoria científica está embasada em alguma teoria filosófica, mesmo que o cientista que a produziu não se dê conta disso. A seguir, vamos identificar a teoria filosófica que suporta a pesquisa de IA, para poder analisar se o projeto de produzir inteligência de modo artificial é viável ou não.

Hubert Dreyfus é um dos maiores críticos da IA. Desde a década de 1960 ele dialoga com os pesquisadores da área, questionando as bases filosóficas que sustentam os projetos da IA. Em 1972, Dreyfus escreveu um livro sobre os limites da IA intitulado "What computers can't do"(DREYFUS, 1975). Esse livro foi lançado em 1975 no Brasil com o título "O que os computadores não podem fazer: Crítica da Razão Artificial", em referência às obras de Immanuel Kant.

Dreyfus afirma que a IA, com o objetivo de reproduzir em computador a inteligência humana, foi construída em cima de pressupostos que foram admitidos sem nenhum tipo de questionamento. Alguns deles acabaram se revelando evidentemente falsos, enquanto outros passaram desapercebidos e são objeto de discussão por alguns pesquisadores até hoje. Dreyfus(DREYFUS, 1975) identifica quatro pressupostos, que ele nomeia como: biológico, psicológico, epistemológico e ontológico.

A seguir, apresentaremos esses quatro pressupostos e a crítica feita a eles.

- **O pressuposto biológico** - O cérebro funciona como um computador digital, ou seja, os neurônios transmitem informações como *flip-flops*.

  Acreditava-se que o cérebro funcionava com seus neurônios transmitindo informações em binário. Isso levou as pessoas a imaginarem que seria uma questão de tempo para os circuitos eletrônicos atingirem a frequência do neurônio e o computador se igualar ao cérebro. O modelo de neurônio mais utilizado atualmente em redes neurais artificiais nos permite descartar esse pressuposto.

- **O pressuposto psicológico** - A mente funciona como um computador digital, processa informações simbólicas tal qual um computador.

  De acordo com aqueles que se utilizam desse pressuposto, existe uma equivalência lógica entre o funcionamento da mente e de um computador. Ou seja, a mente funciona de forma algorítmica, ou ainda, pensamos de forma descontínua através de regras pré-estabelecidas. Na verdade, não existe evidência para tal afirmação. Se o pressuposto psicológico fosse verdadeiro, poderíamos afirmar que ao observar uma cor, nós calculamos os níveis de azul, vermelho e verde ali presentes, o que não parece ser o caso. Isso é equivalente a afirmar que ao arremessar um objeto, ele fará os cálculos para poder descrever uma trajetória curvilínea até atingir o chão.

  Note que essas afirmações invertem a ordem entre o mundo e a teoria que fala a respeito dele. Ou seja, colocam nossos modelos científicos a respeito do mundo como absolutamente corretos e prévios ao fenômeno que ele explica.

  Se a mente funciona como um computador, então deve ser possível produzir teorias psicológicas com base em programas de computador. De acordo com Dreyfus(DREYFUS, 1975), Simon, em 1957, afirmou que algo semelhante ocorreria dentro de dez anos. Infelizmente, o computador não ajudou o avanço da Psicologia como objeto de estudo, somente como ferramenta auxiliar, como ocorreu em inúmeras outras áreas do conhecimento.

- **O pressuposto epistemológico** - Deve existir uma teoria sobre o comportamento humano e a mesma pode ser reproduzida em computador.

  Este pressuposto encontra apoio em duas áreas da ciência: a Física e a Linguística. A Física oferece uma teoria que tenta explicar os fundamentos do comportamento de todos os corpos físicos (inanimados). A ideia então seria estender essa teoria para abarcar os animais e os seres humanos, obtendo-se uma teoria que, de alguma forma, explicasse o comportamento inteligente. Uma tal teoria física do comportamento seria capaz de explicar, a partir das leis da mecânica, seja ela a newtoniana ou a quântica, as causas que fazem as pessoas agirem do modo que agem. Mas, ainda que essa teoria existisse, para que ela fosse útil à Inteligência Artificial, seria necessário realizar uma simulação envolvendo um número incrivelmente grande de objetos. Esse número possivelmente ultrapassaria o limite de Bremermann, que estabelece que nenhum sistema físico é capaz de processar mais de $2x10^{47}$ bits por segundo por grama de sua massa.

Na Linguística, há uma corrente que tenta formalizar a linguagem natural. Tal formalização tornaria possível o entendimento da linguagem natural por parte dos computadores. A ideia é que se a linguagem pode ser reduzida a regras, estas podem ser implementadas em um computador. Mas, apesar de a linguagem possuir regras gramaticais, nós só as utilizamos dentro de um contexto. Mesmo quando as regras são mal aplicadas e há erro, nós conseguimos compreendê-las, graças ao contexto em que a frase foi formulada. Um entusiasta da ideia da formalização da linguagem poderia argumentar que, então, é só uma questão de se formalizar as regras para o contexto, ou seja, um conjunto de regras para se estabelecer em que situações se aplicaria determinada regra. Porém, mesmo que fosse possível fazer regras para um contexto, nós utilizaríamos essas regras também dentro de um contexto, o que tornaria necessário, para o computador, um outro nível de regras. Note que nós não necessitamos desses vários níveis de regras, mas o computador sim. Para a aplicação de regras gramaticais pelo computador, ou caímos em uma regressão infinita de níveis de regras, ou deve existir um nível fundamental que não dependa do contexto e esteja diretamente ligado à realidade. A existência de tal nível é, no mínimo, estranha, já que ela dá a impressão de que o sentido das expressões linguísticas é único, o que obviamente não é verdade. Ademais, esse nível fundamental de regras exigiria que o mundo pudesse ser decomposto em átomos de fatos, que corresponderiam às regras independente de contexto. Essa possibilidade constitui o pressuposto ontológico. Na Seção 2.4.1 veremos que ainda que o computador fosse capaz de aplicar as regras da linguagem natural, o entendimento por parte do computador ainda pode ser questionado.

- **O pressuposto ontológico** - É possível analisar o mundo em termos de dados determinados ou átomos de fatos.

  Um computador funciona de forma discreta, com informações discretas e em etapas discretas, as entradas de seus dados também devem ser discretas. Portanto, se um programa existe para analisar fatos do mundo, estes devem ser divididos de forma discreta. Nesse ponto entra o pressuposto ontológico. Os pesquisadores que o aceitam, acreditam que os fatos do mundo podem ser analisados de forma isolada. Esse é um pressuposto que parece tão autoevidente, que dificilmente é questionado.

  Os fatos do mundo não estão isolados uns dos outros. Não há meios para se analisar fatos do mundo de modo que se consiga distingui-los em fatos mais simples, até se chegar em uma base indivisível. Essa ideia da decomposição atomística, segundo Dreyfus(DREYFUS, 1975), vem desde Platão e ganhou força ao decorrer da história com Galileu, na divisão dos fatos físicos em leis, chegando ao ápice com o primeiro livro de Wittgenstein, que divide o mundo em fatos para fazer uma correspondência com a linguagem. Uma vez que essas ideias foram explicitadas na filosofia, outros filósofos começaram a criticá-las, o próprio Wittgenstein tornou-se seu principal crítico. Ele reformulou sua concepção de linguagem como algo fluido, em que seus conceitos estão em constante processo de ressignificação, dependendo do contexto em que sejam aplicados.

  Uma teoria sobre todos os fatos do mundo pretende apresentar os fatos do mundo em termo de fatos cada vez mais simples. Porém, o que é possível ser feito é uma teoria que se assemelha mais a um dicionário. Um fato é explicado em termos de outros fatos, que

> não necessariamente são mais simples, estes agora devem ser explicados em termos de outros fatos. Por exemplo, ao tentar explicar o conceito de cadeira, teremos que recorrer a outros conceitos bastante complexo, como o de pessoas, o ato de sentar, o número de pernas que possivelmente possa ter uma cadeira, a forma do corpo humano que determina o formato da cadeira e tantos outros. Cada conceito desse não é necessariamente mais simples do que o conceito de cadeira. E cada um deles necessita de outros tantos conceitos complexos para serem explicados. Ou seja, não existe uma necessária simplificação ao se tentar definir fatos do mundo em termos de outros fatos. Nunca chegaremos a fatos atômicos, mais simples que todos os outros, pois eles não existem.

Ao passar dos anos, outros filósofos passaram a criticar as tentativas de inteligência artificial. Vamos analisar três críticas de grande impacto na comunidade científica que tem correspondência com a de Dreyfus. São elas: o argumento do quarto chinês de John Searle (SEARLE, 1980), o *Frame Problem* de Daniel Dennett (DENNETT, 1984) e o *Symbol Grounding Problem* do psicólogo Stevan Harnad (HARNAD, 1990).

Em 1990 Rodney Brooks, pesquisador da área de IA, fez crítica à assunção da hipótese do sistema simbólico físico e apresentou robôs que já não sofreriam do mesmo problema, pois estes eram "*physically grounded*"[4](BROOKS, 1990). No ano seguinte Brooks publicou um artigo (BROOKS, 1991) que é tido por muitos como o pioneiro da nova IA, onde mudou um pouco sua crítica à IA, focando na crítica à representação do mundo. De acordo com ele, esse é o grande entrave para o desenvolvimento de sistemas artificiais realmente inteligentes. Apesar de concordar com Dreyfus nesse ponto, Brooks faz questão de afirmar que seu artigo é fruto de uma observação independente das discussões anteriores acerca da representação.

2.4.1 O Quarto Chinês

Em 1980, John Searle publicou um artigo(SEARLE, 1980) para rebater as afirmações de pesquisadores que diziam ter produzido programas capazes de compreender textos em inglês. O artigo de Searle inicia com um exemplo simples de texto em linguagem natural. No exemplo, uma pessoa vai a um restaurante, pede um prato, depois que o prato chega, a pessoa reclama e vai embora sem pagar. Então, Searle questiona se a pessoa comeu ou não o prato. Ele também dá o exemplo em que a pessoa elogia e dá uma gorjeta, Searle faz o mesmo questionamento a respeito do consumo ou não do prato.Apesar da dificuldade de se automatizar uma resposta para o exemplo, Searle em nenhum momento questiona a capacidade de se produzir algoritmos capazes de responder perguntas em linguagem natural. O ponto central de seus argumentos é a respeito do entendimento do texto por parte da máquina, ou seja, Searle acha que mesmo que um computador seja capaz de interagir com um ser humano através de linguagem natural, ele não será capaz de compreender a linguagem. Searle também não explicita o que seria compreender linguagem, mas se utiliza da noção comum que se tem, sem formalismos.

Para desenvolver seu argumento, o autor propõe um experimento mental em que ele se coloca dentro de um quarto, onde o único contato que tem com o mundo exterior é através

[4] aterrados fisicamente

de pilhas de papel com escritas em uma língua que ele não entende, no caso chinês, e que não é capaz nem de afirmar se tratar de uma língua específica e distinguir de japonês, por exemplo. Alguns desses papéis contém também instruções em inglês, sua língua nativa. No caso do experimento, o idioma chinês faz o papel da linguagem natural e o idioma inglês da linguagem de máquina, já que Searle se põe no lugar da máquina. Além das pilhas com instruções, ele recebe outras pilhas de papel, que contém somente caracteres em chinês e, através das instruções, ele é capaz de escrever símbolos em chinês e devolver para fora do quarto. Para quem está do lado de fora, o quarto recebe uma história e perguntas e devolve respostas, para quem está dentro ele recebe pilhas de símbolos com instruções para combinar outros símbolos e devolvê-los.

Searle pede para que se imagine que os programadores do quarto ficam tão bons em programar, e a pessoa do quarto fica tão boa em combinar os símbolos de forma correta, que os chineses que estão fora do quarto não conseguem distinguir entre as respostas do quarto ou de outra pessoa que tenha chinês como língua nativa. Ou seja, o quarto equivale a um computador que responde perfeitamente perguntas sobre uma história em chinês. Ele ainda pede para que se imagine que também são feitas perguntas para ele em inglês, sobre um texto em inglês, que ele responde tão bem quanto um falante nativo de inglês que ele é. Nesse caso, sem a utilização de regras determinadas.

Agora vem o ponto do argumento de Searle: apesar de conseguir responder perguntas em chinês tão bem quanto um chinês, a pessoa que está dentro do quarto não entende nada de chinês, tudo o que ela faz é manipulação simbólica, em contraste às perguntas em inglês, onde ele não tem regras para seguir e responde tão bem quanto em chinês. Ele entende inglês, mas não entende chinês, sua conclusão é que mesmo que existam computadores que interajam com humanos usando-se de linguagem natural, eles não compreenderão a linguagem.

Searle ainda considera possíveis contra-argumentos, que foram apresentadas a ele em suas palestras, antes da publicação do artigo. Vamos expor alguns dos contra-argumentos e uma resposta a cada um, compatível com as respostas dadas por Searle.

I. Apesar de a pessoa não entender chinês, o sistema como um todo, ou seja, pessoa e sala, compreende chinês.

Suponha que a pessoa que está dentro do quarto internalize todos os elementos do sistema. Suponha que ela decore as regras e não esteja mais dentro de um quarto. Se alguém pergunta para ela algo em chinês, ela ainda pode seguir as regras e responder em chinês. Porém, note que ela, apesar de agora constituir todo o sistema, não compreende chinês, só é capaz de fazer a manipulação simbólica por regras.

II. Considere um robô que anda, consegue ver através de uma câmera, se move, come, bebe, mas com um cérebro computacional, este robô seria capaz de entender.

Esse argumento considera que há alguma necessidade de uma relação com o mundo para haver cognição. Porém, no núcleo do "pensamento" ainda está um computador, ou seja, o processamento simbólico persiste. Como se no caso do experimento do quarto, a pessoa que se encontra lá dentro passasse a receber mais símbolos em chinês. Esse novos símbolos provindos de uma câmera e de sensores externos, mas a pessoa não está ciente disso, ela só recebe os

símbolos. Para quem está dentro do quarto, nada muda.

III. Considere a simulação do cérebro de um chinês nativo, com todas as suas sinapses, disparos de neurônios. Negar que tal simulação entende chinês seria negar que o próprio nativo entende chinês.

Suponha que ao invés de combinar símbolos, a pessoa no quarto opere interruptores, onde cada interruptor corresponde a uma sinapse neural. A pessoa receber um símbolo em chinês, segue as regras programadas, ligando e desligando as sinapses corretas, na ordem correta, e isso gera um símbolo na saída. A pessoa continua sem compreender chinês, nem os interruptores. Como já vimos, nem a combinação da pessoa com os interruptores entende chinês.

IV. Uma combinação dos três anteriores. Um robô, com uma simulação de cérebro e comportamento indistinguível do comportamento humano, seria capaz de entender chinês.

De fato, se fosse um robô com o comportamento e corpo igual ao de um ser humano, certamente todos concordariam que ele seria inteligente. Porém o argumento da IA é de que é possível reproduzir inteligência através de um sistema simbólico físico.

O argumento central de Searle é que sistemas computacionais baseados em manipulação de símbolos não têm intencionalidade, somente a "máquinas" parecidas em todos os aspectos, fisiológicos inclusive, pode ser atribuído intencionalidade.

2.4.2 *Frame Problem*

O *frame problem* foi primeiro descrito por McCarthy e Hayes(MCCARTHY; HAYES, 1969) como o problema de descrever em lógica todos os efeitos que uma ação não acarreta. Por exemplo, encher uma garrafa com água não irá mudar a composição da garrafa, nem sua massa, nem sua cor. Existem muitos efeitos não acarretados por uma ação; conseguir perceber o que é e o que não é relevante em uma ação é inato ao ser humano e crucial para se tomar decisões a respeito do novo cenário em que se está inserido pós ação. Daniel Dennett, filósofo da mente, reavivou esse problema em 1984 (DENNETT, 1984), dando a ele um caráter muito mais abrangente; tornando um problema da Filosofia da Mente, independente da IA. Essencialmente esse é o problema de se estabelecer o conjunto de crenças a respeito do mundo que mudam quando uma ação é executada. De outra forma, quais são todas as consequências de uma ação.

Dennett dá uma descrição de três robôs que "sofrem" do *frame problem*; o primeiro, chamado apenas de "*robot*", não consegue reconhecer os efeitos colaterais de sua ação. O segundo, "*robot-deducer*", perde todo seu tempo deduzindo fatos irrelevantes a respeito de sua futura ação, como por exemplo que a quantidade de água em ml necessária para encher uma garrafa é maior que o número de garrafas que ele deve encher. O terceiro, "*robot-relevant-deducer*", é programado para classificar as suas conclusões como relevante ou irrelevante, o que também toma todo seu tempo. Os dois últimos robôs não chegam a executar ação alguma, tomados por seus cálculos.

Dennett diz que os filósofos nunca conseguiram perceber o *Frame Problem* como

um problema, porque as teorias filosóficas a respeito de como pensamos para agir nunca se aprofundaram muito. Ele compara com um número de mágica onde o mágico serra a mulher ao meio, e a explicação dos filósofos seria que é óbvio que ele não a serra ao meio, ele só faz parecer que sim. A explicação não iria mais longe que isso, porque muita coisa se passa fora do alcance da visão, como na mente. A análise filosófica acaba passando por cima de problemas banais. Na IA, por outro lado, têm-se que fazer tudo a partir do zero, o que torna perceptível problemas como o *Frame Problem* ou outros sobre aprendizado que sejam inatos ao ser humano. Dennett diz que talvez seja inato ao ser humano o *modus ponens*, terceiro excluído ou ainda algum sentido de causalidade.

Segundo o autor, a dificuldade está em fazer algo que não tem conhecimento nenhum sobre o mundo possa ter toda a informação necessária para fazer alterações no mundo. Humanos aprendem pela experiência, mas sistemas de IA são feitos para já saberem muito sobre o mundo *a priori*. Como observamos antes, o problema surge do que Dreyfus chama de pressuposto ontológico(DREYFUS, 1975). Achar que o mundo pode ser analisado com um conjunto de fatos independentes, quando na verdade esses fatos são interdependentes, gera a dificuldade de se computar novamente as dependências.

2.4.3 *Symbol Grounding Problem*

Stevan Harnad faz uma crítica ao modelo simbólico de cognição(HARNAD, 1990), que toma a mente como sendo um sistema puramente simbólico que segue regras definidas. Como o modelo se baseia em um computador, a crítica se estende à Inteligência Artificial. O *Symbol Grounding Problem* tem esse nome porque em um sistema feito só de símbolos, esses ficam "flutuando" uns sobre os outros, sem significado, que é dado por nós, seres humanos, quando "aterramos" os símbolos às nossas experiências sensoriais. Como a IA Tradicional se baseia na hipótese do sistema simbólico físico, seus programas não são capazes de fundamentar o significado, de acordo com Harnad.

Harnad apresenta o primeiro exemplo do *Symbol Grounding Problem* como sendo o quarto chinês de Searle. Logo depois dá mais dois exemplos; um que ele classifica como difícil e outro como impossível. O primeiro exemplo é o de uma pessoa que deve aprender chinês como segunda língua, e cuja única fonte de informação é um dicionário chinês/chinês. A pessoa ficaria dando voltas no dicionário, sem nunca parar, esse é o exemplo difícil. O segundo exemplo é de uma pessoa em uma situação similar, só que dessa vez ela tem que aprender chinês como primeira língua, este é o exemplo impossível. Segundo o autor, os criptologistas só conseguem decifrar línguas antigas por já terem uma linguagem "aterrada"(*grounded*). Mas como aprender uma primeira língua somente através de símbolos que são associados a símbolos? Esse é o *Symbol Grounding Problem*.

Harnad tenta resolver o problema dando um modelo de representação na mente, em que parte das representações estão "aterradas" às percepções sensoriais e outras representações podem ser construídas em cima das que já existem, sendo que as construídas ficam "aterradas" através das outras que serviram como base para elas. Ele dá um exemplo da construção da representação de zebra, que seria uma combinação da representação de cavalo com a representação

de listras. O objetivo de Harnad no seu artigo é criticar o modelo computacional da mente, que Dreyfus percebeu como um pressuposto psicológico. A solução apresentada por Harnad não resolve o problema para a IA, já que o "aterramento" das representações estão ligadas às percepções sensoriais. Como veremos no Capítulo 3, alguns pesquisadores compreenderam as percepções do mundo como necessárias para a inteligência, o que poderia fazer com que a teoria de Harnad também se aplicasse à IA.

3 INTELIGÊNCIA ARTIFICIAL CORPORIFICADA

A Inteligência Artificial tradicional toma como referência a mente humana, e assume, algumas vezes de modo implícito, o seguinte paralelo com os computadores: a mente é o *software* e o cérebro é o *hardware*(RUSSELL; NORVIG, 2010), sendo hipoteticamente possível que uma mesma mente funcione em dois cérebros diferentes, como um programa que é instalado em dois computadores. Nesse modelo, a mente, o corpo e o mundo são três entidades completamente distintas e separadas, que se comunicam por meio de interfaces sintáticas. A mente controla o corpo, que por sua vez está inserido no mundo. Mais uma vez há a possibilidade, dessa vez não tão hipotética, de que uma mente controlando um corpo seja capaz de sobreviver em dois mundos completamente distintos, da mesma maneira que um computador que é mudado de lugar e de tomada. Ainda de acordo com o modelo, o mundo é composto de objetos e fatos que estão previamente definidos, por exemplo, uma garrafa é um objeto do mundo, e ela ser vermelha é um fato.

A Figura 3.1 ilustra as relações das entidades do modelo utilizado pela IA Simbólica, que chamamos de Modelo Mente-Corpo-Mundo. A mente recebe informações sintáticas a respeito do mundo e cria um modelo dele dentro de si mesma. Esse modelo possui todos os detalhes que é possível à mente captar. A mente só consegue entender e raciocinar, de fato, sobre o modelo que ela monta do mundo. Para se locomover, por exemplo, a mente cria um mapa do local onde se encontra. Através desse mapa, ela se guia, calculando distâncias e velocidades. Em seguida a mente dá ordens para o corpo para que se movimente em determinada direção. O corpo só se movimenta a partir das ordens recebidas da mente. O corpo é um instrumento da mente. Ela o utiliza para realizar suas ações e poder interagir com os objetos do mundo. Assim, a mente pode atualizar seu modelo do mundo. Portanto, percebemos que, nesse modelo, a mente é a responsável isolada pela inteligência; todo pensamento e raciocínio acontece dentro da mente. A ideia central do modelo é que a mente trabalha sobre as representações que faz do mundo.

Um exemplo dos problemas que o modelo Mente-Corpo-Mundo sofre: se um robô representacionalista é programado para andar em uma sala com uma mesa no centro, ele terá as dimensões da sala e da mesa armazenadas dentro dele. Quando receber a instrução de ir até a mesa, ele fará um cálculo interno levando em consideração essas dimensões e a sua posição atual, então saberá quantos metros terá de se deslocar, passando a instrução para seu corpo. Se alguém muda a mesa para um canto da sala e dá a instrução para o robô ir até a mesa, ele fará o mesmo cálculo para a posição anterior da mesa, se deslocando até o centro da sala, possivelmente para longe da mesa, a solução seria reprogramar o robô com a mesa no novo local. Porém, ainda no modelo representacionalista, o robô poderia identificar a posição dos objetos, que estão representados nele dentro da sala; assim resolveria o problema de se mudar o lugar da mesa. Ainda assim, digamos que sua representação de mesa tenha quatro pernas e um tampo, se uma das pernas da mesa quebrasse e a colocassem no canto da sala, com a parte da perna quebrada apoiada nas paredes, o robô poderia não reconhecer mais a mesa quebrada como mesa.

Nesse capítulo, falaremos a respeito de um novo paradigma: a IA Corporificada,

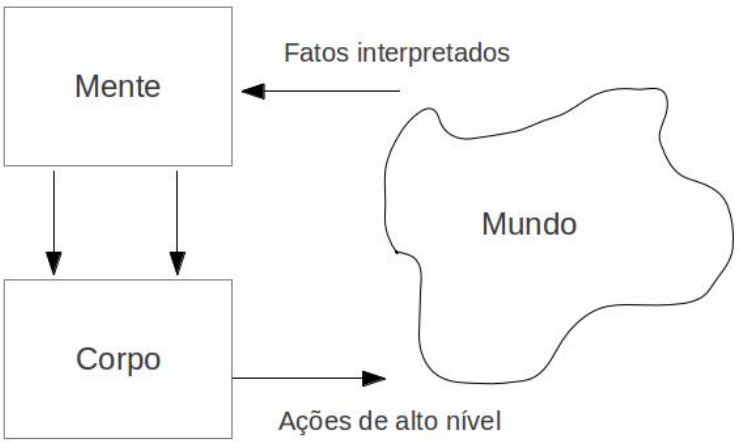

Figura 3.1: Representação do Modelo Mente-Corpo-Mundo
Fonte: Elaborada pelo autor.

que surgiu da negação da representação, colocando o corpo como parte do processo cognitivo. Na Seção 3.1, mostraremos as ideias de Brooks a respeito de uma IA sem representação e sua arquitetura de camadas de inteligência. Na Seção 3.2, abordaremos como o não representacionalismo se deu na área da ciência cognitiva. O desenvolvimento desse paradigma dentro da própria IA será apresentado na Seção 3.3, com alguns experimentos discutidos na Seção 3.4. Finalmente, apresentaremos as críticas ao paradigma na Seção 3.5.

3.1 Inteligência sem representação

No final da década de 1980, a maior parte dos pesquisadores da IA já havia abandonado suas pretensões de reproduzir ou simular o raciocínio humano. A IA tinha se tornado uma área de pesquisa dedicada à resolução de problemas pontuais. A partir de uma divisão das capacidades cognitivas humanas, cada subárea da IA tentava resolver um problema diferente, como o processamento de linguagem natural, o reconhecimento de padrões e o aprendizado automático. Em cada uma dessas áreas o ser humano era o parâmetro, no sentido de que ele oferecia o termo de comparação, mas não necessariamente era a inspiração, por não se impor a restrição de se realizar cada uma dessas tarefas como um ser humano realiza.

Nesse contexto, em 1991, Brooks inicia uma nova vertente na pesquisa em Inteligência Artificial(BROOKS, 1991), interessada novamente em perseguir a meta de inteligência humana. Ele lança a ideia de que a inteligência deve ser construída incrementalmente. Brooks entende que existem vários níveis de inteligência desde de uma forma de vida simples como um inseto até o ser humano. Tomando como base a teoria evolucionária, Brooks acredita que a inteligência das formas de vida mais simples é a base para o desenvolvimento das capacidades cognitivas mais complexas.

Brooks se inspirou no funcionamento do cérebro, que se subdivide em áreas relativamente independentes, e criou o que chamou de *subsumption architecture*, ou arquitetura de subsunção. Tal arquitetura é composta por partes que, como o nome sugere, vão sendo incluídas no sistema já existente. Essas partes são como módulos, onde cada um é independente dos demais e só é incluído em um sistema já testado e robusto. O novo sistema, composto do sistema antigo mais o módulo incluído, é então testado e avaliado, podendo o módulo ser modificado no caso de possíveis conflitos com os módulos mais antigos. Quando o novo sistema está robusto o suficiente, ele está pronto para receber um possível novo módulo. Por exemplo, em geral o primeiro módulo dos testes de Brooks faz como que o sistema, no caso o robô, evite bater nos obstáculos. Essa é a única tarefa do módulo, ele é testado e, uma vez considerado robusto, o sistema é considerado pronto para receber um novo módulo. Um segundo módulo comum nos robôs de Brooks faz o robô andar aleatoriamente quando não está ocupado desviando de objetos. O robô com os dois módulos é capaz de andar por aí sem colidir com outros objetos, mas isso não é feito de forma centralizada. Cada módulo tem seu próprio processador e eles disputam o controle dos motores do robô, podendo haver uma ordem de prioridade entre eles. No caso, é mais importante não bater em outros objetos do que continuar vagando.

Outra diretriz para o desenvolvimentos dos robôs de Brooks consiste em usar pouco processamento a respeito das decisões que o robô toma para agir, isto é, pouco processamento em cada camada. Deve haver, também, uma interação mais próxima entre os sensores e os motores do robô, sem que tudo seja algoritmicamente computado por uma entidade central. Essa relação próxima de sensores e motores recebeu o nome de laços sensório-motores (*sensorimotor loops*). Estes compõem um sistema de retroalimentação direta, ou seja, aquilo que é captado por um sensor influencia os motores sem ter que passar por uma unidade central. Em resumo, a arquitetura desenvolvida por Brooks o permite utilizar unidades de processamento descentralizadas, que requerem baixo poder de processamento e promovem uma interação mais direta entre os sensores e motores.

A grande contribuição de Brooks foi ter proposto um modelo para a inteligência que não utiliza uma representação do mundo. Brooks afirma que em um nível de inteligência simples, "representations and models of the world simply get in the way. It turns out to be better to use the world as its own model."[1](BROOKS, 1991). Por esse motivo, Brooks trabalha com robôs e não com softwares, para que estes possam interagir com o mundo real. Segundo ele, treinar robôs em mundos mais simples antes de soltá-los no mundo real seria prejudicial por condicioná-los a reagir a formas que no mundo real não existiriam. A arquitetura de subsunção foi desenvolvida por Brooks com o propósito de incorporar suas metas de não representação, através da ligação feita com o laço sensório-motor. Note que esse forma um ciclo que passa por fora do agente. Analisando esse ciclo começando pelos sensores, esses captam informações do mundo, enviam para as unidades de processamento, que influenciam os motores. O ciclo se fecha quando a ação dos motores altera o ponto de vista da percepção do robô sobre o mundo, pela mudança de posições dos sensores, ou do próprio robô. Esse ciclo, passando por fora do robô, mostra que nesse modelo o ambiente também influencia no processo cognitivo.

[1]Representações e modelos do mundo simplesmente atrapalham. Acontece que é melhor usar o mundo como seu próprio modelo.

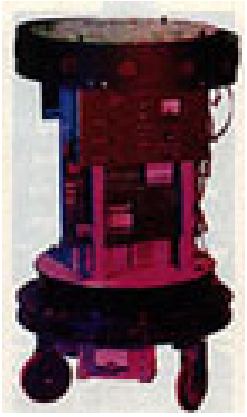

Figura 3.2: Allen
Fonte: http://alife.tuke.sk/kapitola/1153/index.html

Note que a falta de uma unidade central nos robôs de Brooks impede a existência de uma representação nos moldes da IA Tradicional. Para haver representação nesses moldes, deveria haver uma representação em cada módulo, o que não é o caso. Além disso, a interação entre sensores e motores, do modo como é feita, é incompatível com uma representação interna do mundo. Porém, ainda podemos observar nas implementações de Brooks a pré-interpretação do mundo por parte do programador. Ou seja, ao projetar o robô, Brooks diz qual o propósito do robô; com isso o constrói de modo a ter aquele comportamento específico. Isso ainda caracteriza a presença da anterioridade, que falamos no Capítulo 1. Na Seção 3.5 falaremos um poucos mais sobre isso. A seguir, apresentamos alguns robôs construídos por Brooks.

3.1.1 Os Robôs de Brooks

Nesta seção, detalhamos algumas implementações da arquitetura de subsunção de Brooks em seus robôs descritos em (BROOKS, 1990). As partes da arquitetura que chamamos de "módulos" para explicar o funcionamento da arquitetura, Brooks as chama de "camadas".

Allen, Tom e Jerry

Os três possuem três camadas em suas arquiteturas, sendo que Tom e Jerry são totalmente idênticos e Allen (Figura 3.2) se diferencia na implementação de sua terceira camada. A primeira camada tem um algoritmo que faz o robô desviar de obstáculos, inclusive dos que se aproximam enquanto ele está parado. A segunda camada faz com que o robô ande aleatoriamente. Já a terceira, em Allen faz com que o robô procure através dos sonares, por locais distantes e vá naquela direção. Em Tom e Jerry, é implementada para que os robôs sigam algum objeto. Um fato interessante a respeito dos três robôs é que a terceira camada suprimiu o comportamento da segunda sem prévia programação.

Hebert

Hebert (Figura 3.3) foi programado para andar pelos escritórios e coletar latas de refrigerante vazias deixadas em algum canto ou que sejam colocadas em sua "mão". Ele demonstrou capacidade de evitar objetos, seguir paredes, reconhecer as latas e um conjunto de 15

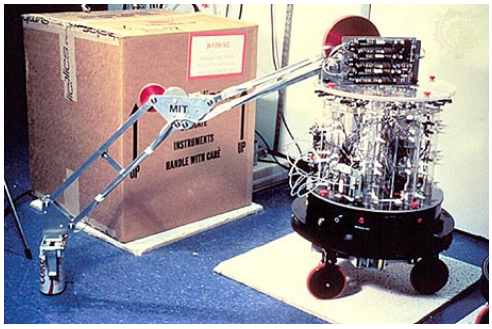

Figura 3.3: Hebert
Fonte: http://www.britannica.com/EBchecked/media/55524/Herbert-the-robot-1987-Designed-by-Rodney-Brooks-and-affectionately

Figura 3.4: Genghis
Fonte: http://alife.tuke.sk/kapitola/1153/index.html

comportamentos que levavam o braço a procurar, localizar e pegar a lata. Hebert não possui comunicação interna entre seus módulos geradores de comportamento.

Genghis

Genghis (Figura 3.4) possui camadas que permitem que ele fique de pé, ande sem usar sensores, é capaz de passar por terrenos acidentados, passar suas pernas sobre obstáculos, inibir a passagem por terrenos acidentados, detectar pessoas e andar somente quando elas estiverem presentes.

3.2 Cognição Corporificada

A ciência cognitiva é uma área multidisciplinar da qual a IA faz parte, juntamente com a psicologia, a neurociência, a filosofia, entre outras. As críticas à IA Tradicional repercutiram na ciência cognitiva como críticas ao modelo que tomava como base justamente o computador; o modelo computacional da mente do qual falamos na Seção 2.4. Esse modelo tem dificuldades de explicar, por exemplo, como os símbolos dentro do cérebro adquirem significado. Outra dificuldade do modelo é a ideia de que um ser inteligente tem uma representação do mundo dentro de si, como falamos no início do capítulo, o que leva à ideia de um homúnculo controlando o corpo com base nessa representação. Mas, a cognição para esse homúnculo só pode ocorrer de duas formas: ou ele acessa diretamente as representações que chegam até ele,

ou ele mesmo, homúnculo, também produz representações para poder raciocinar. A primeira hipótese nos leva à ideia de que o homúnculo é desnecessário. Pois, se é possível ter acesso direto ao objeto da cognição, o próprio ser inteligente poderia fazê-lo. A segunda hipótese nos levaria a uma regressão infinita. Já que se o homúnculo produz representações internas, deveria haver algo dentro dele que raciocina sobre as suas representações.

Surgiu então, no âmbito da ciência cognitiva, um modelo que se contrapõe ao modelo computacional da mente, a chamada cognição corporificada (*embodied cognition*). A principal característica da cognição corporificada é, como o nome sugere, a observação de que o corpo também participa do processo da cognição. Essa característica torna a área um pouco vasta e imprecisa. De acordo com Wilson e Foglia, a cognição corporificada se caracteriza por aceitar o que chamam de "Embodiment Thesis", que diz:

> Many features of cognition are embodied in that they are deeply dependent upon characteristics of the physical body of an agent, such that the agent's beyond-the-brain body plays a significant causal role, or a physically constitutive role, in that agent's cognitive processing[2](WILSON; FOGLIA, 2011).

Shapiro em (SHAPIRO, 2007) destaca três subáreas da cognição corporificada que se relacionam. A primeira, que acredita que parte do processo cognitivo emerge do corpo. A segunda, que afirma que o conteúdo cognitivo depende da forma do corpo que contém o cérebro. A terceira, que diz que o processo de cognição não se restringe ao sistema nervoso, mas se estende até o ambiente em que o corpo está inserido.

Ainda em (SHAPIRO, 2007), o autor dá um exemplo para a primeira subárea de como o corpo pode influenciar a cognição. Ele afirma que a distância entre as orelhas afeta a acuidade auditiva, quanto maior a distância melhor é a audição. Porém, a densidade da matéria entre as orelhas também é importante, pois os sons diferem quando passam por meios diferentes.

Um trabalho muito importante para a cognição corporificada é o de Lakoff e Johnson(LAKOFF; JOHNSON, 1980), que é um exemplo para a segunda subárea. Esse trabalho diz que a linguagem humana é composta por muitas metáforas, e que sempre aprendemos coisas novas a partir de relações com o que já sabemos. Um exemplo de Lakoff e Johnson é a metáfora do amor e viagem. O amor pode ser entendido como uma metáfora para uma viagem: ele tem um início, pode não ter um fim, pode levar a lugares inesperados, pode ser difícil em alguns momentos, mas recompensador em outros. Mas, se o que aprendemos tem relações com o que já sabemos, deve existir uma base para não cairmos em outra regressão infinita. De acordo com Lakoff e Johnson(LAKOFF; JOHNSON, 1980), a base é dada pela forma do nosso corpo, de onde tiramos ideias como cima, baixo, frente, trás, perto, longe etc. Os conceitos de cima e baixo seriam desenvolvidos de forma inata a partir da necessidade de o ser humano ficar em pé e ter que manter ou mudar a orientação cima-baixo, por exemplo. Shapiro diz que Lakoff e Johnson oferecem outras explicações para o restante das ideias citadas. Em (SHAPIRO, 2007),

[2]Muitos aspectos da cognição estão incorporados de modo que são profundamente dependentes de características do corpo físico de um agente, tal que o corpo além-do-cérebro do agente tem um papel causal significante, ou um papel fisicamente constitutivo, no processamento cognitivo do agente.

o autor cita pesquisas de Arthur Glenberg (GLENBERG, 1997)(GLENBERG; ROBERTSON, 2000) que apontam uma maior dificuldade para seres humanos entenderem conceitos que não são adequados com o corpo humano.

Para a terceira subárea, Shapiro cita o trabalho de Wilson(WILSON, 2004), que afirma que ao se utilizar papel e caneta para auxiliar na resolução de um problema matemático, estes objetos são uma extensão do aparato cognitivo; eles estão integrados no ato cognitivo de tal forma que não existe razão para distingui-los do resto do sistema cognitivo na resolução do problema. O autor ainda cita Clark(CLARK, 1998), que diz que o ser humano organiza o seu ambiente de tal forma que facilite seus processos mentais, como por exemplo deixar as chaves sempre perto da porta, para não ter que lembrar onde as deixou, ou organizar arquivos em ordem alfabética para facilitar buscas. Assim outros seres humanos também teriam seus processos cognitivos facilitados,se a organização é compartilhada por eles como no caso da ordem alfabética.

3.3 Inteligência Artificial Corporificada

Os trabalhos de Brooks colocaram em oposição a representação e o par percepção-ação. Através da aproximação entre a percepção e a ação de seus robôs, Brooks procurou eliminar a utilização de representação por parte dos mesmos. Na concepção tradicional da IA, a representação tem um papel crucial. Ela é quem possibilita a mente realizar cálculos a respeito dos fatos do mundo e decidir como controlar o corpo. Toda interação entre a mente, o corpo e o mundo é realizada através de interfaces sintáticas que "traduzem" a informação de um lado para outro. O modo como ocorre tal tradução é previamente determinado pelo programador, que deve perceber o que há de relevante de um lado que deve passar para o outro. Na concepção de Brooks, a divisão entre as partes do sistema se dá de outra forma. O que existe em seu trabalho é o agente e o mundo. A mente e o corpo não são elementos dissociados, eles se confundem dentro do agente. Mesmo entre o agente e o mundo a relação se dá de forma mais direta, as interfaces sintáticas não existem mais. O modo como o mundo afeta o agente e *vice versa* é através dos sensores e motores do agente. Quando há uma mudança no mundo, ela só é percebida pelo agente se afetar algum dos seus sensores. Os sensores geram um sinal para dentro do agente que ativa seus motores de alguma forma. Uma vez que os motores foram ativados, a posição relativa do agente, ou de um de seus membros, se modifica. Assim a percepção do mundo pelo agente não é mais a mesma, o que afeta novamente seus sensores, fechando um ciclo. A esse ciclo dá-se o nome de laço sensório-motor. A parte da IA que leva em conta o corpo no processo cognitivo, visando aproximar o par percepção-ação através dos laços sensório-motores, é chamada de IA Corporificada.

A área da IA Corporificada surgiu e começou a se desenvolver sem nenhum tipo de critério. Os cientistas rotulavam seus próprios trabalhos como pertencentes à área. Algumas vezes, bastava uma característica, como por exemplo o fechamento do ciclo entre percepção e ação para que o autor do trabalho o classificasse como sendo da IA Corporificada, mesmo que o trabalho se utilizasse de representação do mundo dentro do agente. Essa falta de sistematização da área levou Pfeifer a propor que a comunidade de IA Corporificada estabelecesse princípios

Rótulo	Nome	Descrição
P1	Metodologia Sintética	Entender através da construção.
P2	Emergência	Sistemas devem ser projetados para emergência.
P3	Tendência a diversidade	Problema de troca entre explorar o que é dado e gerar diversidade resolvido de formas interessantes.
P4	Perspectivas de tempo	Três perspectivas: Aqui e agora, ontogenética, filogenética.
P5	Quadro de referência	Três aspectos devem ser distinguidos: perspectiva, comportamento vs. mecanismos, complexidade.

Tabela 3.1: Princípios de Procedimento de Projeto
Fonte:(PFEIFER; IIDA; BONGARD, 2005)

norteadores para caracterizar melhor a área. Em 2005, juntamente com Iida e Bongard, Pfeifer publicou sua sugestão de como deveriam ser esses princípios (PFEIFER; IIDA; BONGARD, 2005). Eles foram desenvolvidos e modificados durante anos (PFEIFER; IIDA, 2003), mas nem por isso os autores alegam que os princípios são definitivos. Pfeifer *et al.* (PFEIFER; IIDA; BONGARD, 2005) afirmam que construíram os princípios a partir das ideias de comportamento adaptativo (*adaptive behavior*) e da ciência cognitiva. Comportamento adaptativo é um tipo de comportamento que permite que o indivíduo modifique seu comportamento de modo a se adaptar às restrições que o ambiente lhe impõe. A ciência cognitiva influenciou Pfeifer *et al.* através das ideias de que o corpo participa do processo cognitivo, como discutimos na Seção 3.2. O objetivo de Pfeifer ao propor os princípios norteadores da área era lançar alguns princípios para que a comunidade de IA Corporificada discutisse e que os princípios acabassem evoluindo através de uma discussão dos pesquisadores, que modificariam os princípios existentes e sugeririam novos princípios. Os autores deixam em aberto a possibilidade de existirem mais princípios a serem desenvolvidos pela comunidade de IA.

Os princípios de Pfeifer *et al.* estão divididos em dois tipos: os princípios de procedimento de projeto ("*design procedure principles*"), que eles chamam de princípios P, e os princípios de projeto do agente ("*agent design principles*"), que os autores chamam de princípios A. Os princípios P estão ligado à maneira que se deve proceder em um projeto de IA Corporificada e o modo que se deve analisá-lo. Os princípios A dizem respeito ao projeto do agente em si, seja ele um robô ou uma simulação. As Tabelas 3.1 e 3.2 reproduzem os princípios P e A, respectivamente, descritos por Pfeifer, Iida e Bongard em (PFEIFER; IIDA; BONGARD, 2005), que iremos detalhar a seguir.

O princípio P1 trata da compreensão do sistema através da sua construção, ou seja, esse princípio guia o projetista a se preocupar em como construir um sistema de modo que ele possa ajudar na compreensão do fenômeno que reproduz. Por exemplo, um robô que reproduz o andar de um inseto deve ser construído de modo a ajudar na explicação da locomoção do inseto, ou uma simulação de inteligência humana deve ajudar a explicar a mesma. Segundo Froese e Ziemke, o princípio P1 ajuda-nos a compreender o fenômeno da vida e da mente. Isso fica claro se pensarmos no projeto de um agente com características de ser vivo e da mente humana feito a partir do princípio P1. O princípio P2 envolve o conceito de emergência. A emergência é um fenômeno que ocorre em sistemas cuja interação de suas partes gera, ou faz emergir, um comportamento global do sistema não redutível às partes. No caso do princípio P2,

Rótulo	Nome	Descrição
A1	Três constituintes	Nicho ecológico, tarefas e o agente sempre deve ser levado em conta.
A2	Agente completo	Agentes corporificados, autônomos, autossuficientes, situados.
A3	Processos paralelos, folgadamente acoplados	Processos paralelos, assíncronos, parcialmente autônomos, largamente acoplados através de interação com o ambiente.
A4	Coordenação sensório-motora	Comportamento sensório-motor coordenado com respeito ao alvo; estímulo sensorial auto-gerado.
A5	Projeto barato	Exploração de nicho e interação; parcimônia.
A6	Redundância	Sobreposição parcial de funcionalidades baseadas em processos físicos diferentes.
A7	Equilíbrio ecológico	Equilíbrio na complexidade dos sistemas sensorial, motor e neural: distribuição de tarefas entre morfologia, materiais e controle.
A8	Valor	Forças motrizes; mecanismos de desenvolvimento; auto-organização.

Tabela 3.2: Princípios de Projeto do Agente
Fonte:(PFEIFER; IIDA; BONGARD, 2005)

a emergência está presente no sentido da não pré-programação. Para Pfeifer *et al.*, o comportamento do agente deve emergir e não ser programado. Os autores afirmam que a emergência é um fenômeno gradual e que quanto menos for programado e mais emergir, melhor. Emergência pode dar ao agente um grau de adaptação, já que de situações diferentes podem emergir comportamentos diferentes.

O princípio P3 fala sobre a variedade de comportamento que o agente deve apresentar em situações similares. A ideia é que o comportamento de um ser vivo não é algorítmico, seu comportamento deve variar pelo menos um pouco quando se encontra em uma mesma situação pela qual já passou. Um ser humano, por exemplo, não consegue repetir exatamente o mesmo movimento duas vezes, a não ser que treine muito. O princípio P4 diz que uma explicação a respeito do comportamento do agente tem que considerar que ele está embutido em três escalas de tempo: o presente imediato, o desenvolvimento do próprio agente durante a sua vida, e a evolução da espécie ou do seu nicho. Essas três escalas de tempo não são separadas, elas só aparecem assim para efeitos de entendimento. O último dos princípios P trata do esquema de referência ao se analisar um agente. Primeiro deve-se saber em que perspectiva ele está sendo analisado, se da perspectiva do próprio agente, de um observador externo ou do projetista. Se sempre nos posicionarmos como um observador externo, por exemplo, o argumento do quarto chinês torna-se nulo. Outro aspecto deste princípio é que ele lembra o princípio P2 ao dizer que o comportamento não é redutível a mecanismos internos. Por último, o princípio P5 também fala que um aparente comportamento complexo não implica complexidade do mecanismo, a complexidade do ambiente que o agente habita também é fundamental no surgimento de comportamentos complexos.

Vamos detalhar agora os princípios do agente. O princípio A1 diz que o agente não deve ser projetado isolado. Ao se projetar um agente corporificado, deve-se levar em consideração as tarefas que se deseja que ele execute, o ambiente e o próprio agente. O princípio

A2 afirma que os agentes devem ser independentes com relação a outros agentes, devem ser autossuficientes e projetados como um sistema físico e obter informações do ambiente através de seus sensores. O princípio A3 diz que a cognição é resultado da interação de vários processos paralelos, assíncronos e não rigidamente acoplados. O princípio A4 diz que todo comportamento inteligente deve ser concebido como coordenação sensório-motora. O agente deve estruturar seus sensores através de uma interação eficiente com o ambiente. Já presente nos robôs de Brooks, esse princípio é o que mais diretamente se opõe ao representacionalismo da IA Simbólica.

O Princípio A5 diz que se deve aproveitar as características físicas do ambiente do agente. Como, por exemplo, robôs que utilizam rodas por seu ambiente ser feito só de áreas planas. Esse princípio está relacionado com os princípios P3 e A1, pois ambos sugerem o entendimento e construção integrado do agente e ambiente. A6 fala da importância de subsistemas do agente se sobreporem em respeito a alguma funcionalidade com intuito de dar mais robustez ao agente. Não é uma questão de ter dois subsistemas iguais, mas subsistemas distintos que digam respeito a uma mesma funcionalidade. Por exemplo, o sistema visual e tátil ajudam-nos a compor nossa noção espacial. A7 é um princípio com duas partes: a primeira diz que a complexidade do ambiente deve ter uma equivalência na complexidade do agente; mundos mais complexos requerem sistemas sensoriais, motores e neurais mais complexos. A segunda diz que dado um ambiente, deve-se projetar a morfologia e o material para que se facilite o controle; o objetivo é que se consiga acoplar a dinâmica física com a dinâmica do controle. O último princípio, A8, diz que o agente deve ter um sistema de valores para distinguir ações boas de ações ruins, podendo modificar e evoluir seus comportamentos.

Podemos observar que os princípios propostos por Pfeifer *et al.* possuem algumas redundâncias e a falta de um esquema conceitual de mais alto nível que organize os princípios, para além da simples divisão em dois tipos. Como dissemos, Pfeifer lançou a ideia dos princípios para que, com a contribuição da comunidade de IA Corporificada, a área passasse a ser mais organizada. Porém, a falta de desenvolvimento desses princípios de forma explícita ou implícita evidencia que a área ainda não está madura.

3.4 Experimentos em IA Corporificada

Nessa seção, iremos relatar três trabalhos de IA corporificada que ilustram o estabelecimento do laço percepção-ação e como ele se adapta a modificações no ambiente.

3.4.1 Di Paolo

Em (DI PAOLO, 2000) Di Paolo utiliza uma simulação de robôs para testar as ideias de Ashby sobre homeostase(ASHBY, 1960) *apud* (DI PAOLO, 2000). De acordo com essa teoria, os seres vivos possuem "variáveis essenciais" que o corpo tem que manter estáveis para a sua sobrevivência. Quando há estabilidade o ser vivo está adaptado, e quando há instabilidade o ser vivo precisa realizar alguma ação de modo a reequilibrar suas variáveis essenciais. Por

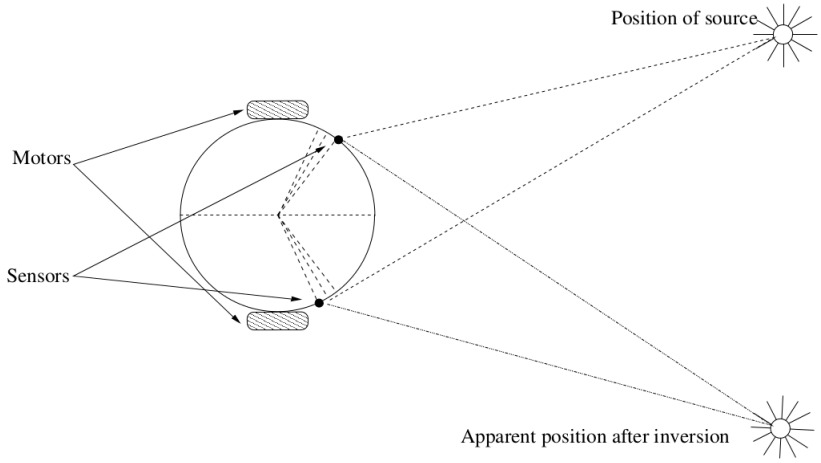

Figura 3.5: Robô de Di Paolo treinado para fazer fototaxia
Fonte:(DI PAOLO, 2000).

exemplo, em animais, a fome é um indício de que uma variável essencial está instável, para reequilibrar suas variáveis o animal precisa comer.

Di Paolo recria em robôs simulados um experimento já aplicado em animais e seres humanos, em que estes usam óculos que invertem a imagem que chega ao sujeito da experiência. Há casos de experimentos que fazem a inversão horizontal e casos em que fazem a inversão vertical. Em seres humanos o experimento mostrou que o indivíduo usando os óculos, em um primeiro momento, não consegue distinguir objetos no seu campo de visão. Somente quando começa a se locomover e interagir também com o tato, esse indivíduo volta a enxergar como antes de vestir os óculos. No experimento de Di Paolo, ele cria robôs simulados e pretende que, através da aplicação do conceito de homeostase, os robôs consigam se orientar depois de terem seus sensores de luz invertidos.

Os robôs são circulares com um motor em cada lado, que podem ir para frente e para trás de forma independente. Os robôs têm também dois sensores de luz conforme a Figura 3.5, podendo variar conforme as linhas pontilhadas. O robôs são equipados, ainda, com redes neurais de oito neurônios, em que cada motor e cada sensor têm dois neurônios para controlá-los e os neurônios são totalmente conectados entre si. O autor treina os robôs para se locomoverem em direção a um ponto de luz que aparece no seu ambiente; esse tipo de locomoção é chamada de fototaxia. Durante o treinamento a fonte de luz aparece por um longo período de tempo; depois muda de lugar, reaparecendo dentro do campo de visão do robô, mas um pouco mais distante. Uma vez treinados, ou seja, quando já conseguem se aproximar da luz toda vez que ela muda de lugar, os robôs têm os seus sensores invertidos.

Metade dos robôs treinados atingiu a estabilidade para fototaxia, e desses, a metade conseguiu se adaptar quando tiveram seus sensores invertidos. Além disso, foi constatado que quanto maior o tempo que o robô passa antes de ter os sensores invertidos, maior é o tempo necessário para se adaptar depois da inversão. Para Di Paolo, essa relação entre o tempo de adaptação e o tempo pré-inversão demonstra que a rede neural tende a se "enrijecer", no sentido de que fica mais difícil de se modificar, quanto maior for seu treinamento. Di Paolo conclui que,

a capacidade de suas redes neurais de fazer com que o robô se readapte para realizar fototaxia mesmo com os sensores invertidos, é um passo importante na direção da implementação do conceito de homeostase.

3.4.2 Wood e Di Paolo

Em (WOOD; DI PAOLO, 2007) Wood e Di Paolo também investigam a homeostase através da simulação de robôs. A ideia é repetir um experimento feito com crianças entre sete e doze meses de idade realizado por Piaget. Nesse experimento uma criança fica sentada a uma mesa que tem dois buracos, A e B; um cientista chama a atenção da criança com um brinquedo e o coloca no buraco A, e pouco tempo depois é permitido à criança pegar o brinquedo. Esse procedimento é repetido algumas vezes, e em seguida o cientista coloca o brinquedo no buraco B, esperando o mesmo tempo de antes para permitir que a criança pegue o brinquedo. No experimento a criança sempre tenta pegar o brinquedo no buraco A, mesmo tendo visto que ele foi colocado no buraco B.

Os autores repetem o experimento com robôs para testar se modelos de simulação minimal podem auxiliar, de alguma forma, a explicação de fenômenos complexos do desenvolvimento humano. Um robô circular com dois motores diametralmente opostos e dois sensores acústicos posicionados simetricamente a 45°dos motores é utilizado. O robô pode mover um pouco seus sensores em volta do seu corpo, podendo controlar sua "atenção" auditiva. Os robôs podem ir para frente e para trás, seus corpos são rígidos, pequenos e de pouca massa. Usa uma rede neural recorrente de tempo contínuo com nove nós, um para cada sensor, um para cada motor, uma rede completamente conectada de quatro nós e o último neurônio para controlar o movimento dos sensores em volta do corpo.

O experimento é feito auditivamente com duas fontes sonoras colocadas simetricamente em relação a posição inicial do robô. Cada tentativa é dividida em três fases, conforme ilustra a Figura 3.6: na primeira fase uma das fontes emite som, na segunda fase há a espera, e, como no experimento original, na terceira fase o robô deve se aproximar da fonte que estava emitindo som na primeira fase. Em todas as fases é permitido ao robô rotacionar os sensores. Foram feitos dois experimentos, que se diferem no modo como os pesos das conexões entre neurônios são calculados conforme descreveremos.

Experimento 1 Aqui, os pesos das conexões entre os neurônios são configurados por um mecanismo baseado em aprendizado Hebbiano. Além disso, utiliza-se um algoritmo genético para evoluir a rede, sem a restrição de homeostase. A primeira tentativa produziu robôs com dificuldade de cumprir a tarefa e com comportamento estereotipado. Então, o experimento sofreu uma leve modificação para que na fase de aproximação ambas as fontes emitissem um som alto. Desse modo, o robô não perde a referência do local em que se encontram as fontes sonoras. Com a modificação, o experimento produziu robôs que em dez de doze tentativas acertaram qual a fonte que emitiu som na primeira fase, com trajetórias menos estereotipadas. Poucos robôs desse experimento tiveram o padrão perseverante que as crianças apresentaram no experimento original, ou seja, poucos robôs erraram a fonte quando elas foram trocadas.

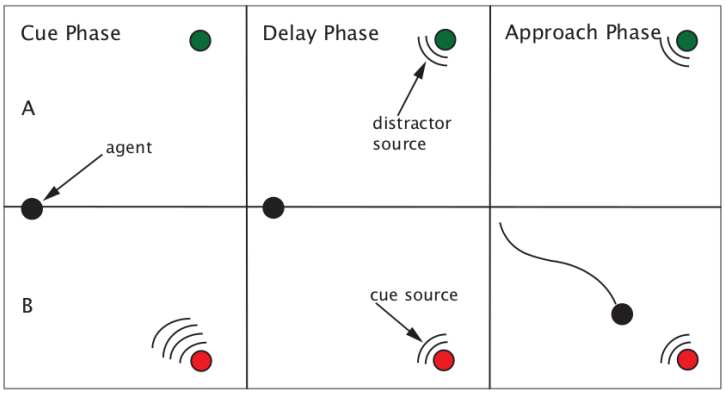

Figura 3.6: Fases do Experimento de Wood e Di Paolo
Fonte:(WOOD; DI PAOLO, 2007)

Experimento 2 Nesse experimento há uso da teoria da homeostase. Aqui, a variável essencial é a frequência com que o neurônio "dispara", ou seja, emite sinais de saída. Esta frequência determinará os pesos dos sinais de entrada no mesmo neurônio. O experimento 2 gerou robôs com menos erros, a maioria deles presente no período de troca da fonte emissora. O que caracteriza, segundo os autores, que os erros deste experimento foram perseverantes. Foi ainda observado que os erros durante a troca diminuem com o aumento de tentativas.

Os autores se utilizam da simulação de aspectos da cognição na tentativa de auxiliar na explicação de fenômenos da cognição humana. Porém, o experimento demonstra que o robô resolve de maneira simples o problema. Para haver semelhança com o experimento original, é necessário levar mais aspectos da cognição humana em consideração, no caso a homeostase. Certamente o experimento não reproduz a forma humana de resolver o problema e não se sabe o quanto de outros aspectos falta. Então, o melhor a se fazer é resolver o problema da implementação de inteligência artificial primeiro para depois utilizá-la para traçar paralelos com a cognição humana.

3.4.3 Izquierdo e Harvey

Em (IZQUIERDO; HARVEY, 2007) Izquierdo e Harvey investigam a possibilidade de um robô com plasticidade reaprender. Baseados em um experimento feito com nemátodos, em que o verme faz uma associação de temperatura e comida, os autores reproduzem o experimento utilizando robôs. Nesse experimento, os robôs têm que fazer uma associação entre a temperatura e a qualidade da comida em dois tipos de ambiente e reaprender, ou seja, modificar sua preferência por temperatura quando necessário.

O ambiente do robô é composto por uma arena de duas dimensões com gradiente térmico em uma das dimensões e contém comidas nutritivas e comidas venenosas. Há dois tipos de ambientes: os $\top$ e os $\bot$. Nos ambientes tipo $\top$ a comida nutritiva se encontra na região quente do ambiente. Já nos ambientes tipo $\bot$, a comida nutritiva se encontra na região fria.

O robô é constituído de corpo circular, com dois motores diametralmente opostos e dois sensores, um de temperatura e um sensor de comida. O robô só pode se mover para

Estágio	1	2	3	4	5
φ	$[0,\pi]$	$[0,2\pi)$	$[0,2\pi)$	$[0,2\pi)$	$[0,2\pi)$
k	1	1	1	2	5
p	1	1	5	[1,5]	[1,5]

Tabela 3.3: Parâmetros dos estágios de complexidade do experimento
Fonte:(IZQUIERDO; HARVEY, 2007)

frente e virar. A comida só pode ser percebida se o agente estiver muito próximo. Os testes foram feitos com robôs utilizando redes neurais de três, quatro e cinco neurônios. Durante o experimento o robô pode ser mudado de lugar abruptamente, o ambiente pode mudar de tipo e a posição angular inicial do robô também pode se modificar. Os autores estabeleceram, conforme a Tabela 3.3, cinco estágios de complexidade do ambiente. Onde φ representa o a variação do ângulo inicial, k é o número de mudanças na posição do robô e p a quantidade de mudanças do tipo de ambiente. Um robô só passa para o estágio seguinte se tiver sido bem sucedido no estágio em que se encontra. Ou seja, se conseguir buscar comida nutritiva mesmo com as mudanças do estágio. Nenhum robô com três neurônios passou do estágio dois. A maioria dos robôs com quatro neurônios foi até o estágios dois, mas muitos foram até o estágio cinco. Entra os robôs com cinco neurônios, a grande maioria foi até o estágio cinco. Os autores resolveram, então, analisar os robôs com quatro neurônios por terem sido os mais simples a chegarem ao último estágio.

No estágio cinco, os robôs têm mudança da posição angular inicial no intervalo $[0,2\pi)$, são movidos abruptamente cinco vezes e ocorrem entre uma e cinco mudanças no tipo de ambiente. Na análise dos agente de quatro neurônios, os autores mostram que a função objetivo do algoritmo genético utilizado para treinar os robôs cai drasticamente nas primeiras mudanças de ambiente, porém quase não se altera nas últimas.

A performance dos melhores robôs foi testada em mais dez experiências de avaliação com dez mudanças de ambiente cada. Os agentes obtiveram sucesso na maior parte das vezes. Primeiro, o robô procura comida navegando gradiente abaixo, mas muda de direção antes de atingir a região onde normalmente se encontra a comida; ele ainda não sabe em que tipo de ambiente se encontra. Quando é movido abruptamente depois de encontrar comida, ele vai diretamente onde a comida estava anteriormente. Quando há mudança no tipo de ambiente, o agente só navega em direção contrária à comida na primeira tentativa, depois vai diretamente para o lado correto. Também foi testado um ambiente com comida nutritiva em ambos os lados; nesse caso, o agente prefere ir sempre para onde encontrou comida na última tentativa.

A interpretação feita pelos autores do trabalho parece confundir a percepção do agente com a percepção de um observador externo. Isso se evidencia ao utilizarem conceitos subjetivos para descrever ações do robô, falando da preferência do mesmo. Quanto a reaprendizagem, esse conceito está ligado ao fato de se modificar a aprendizagem que se tinha a respeito de algo, sendo que o que tinha sido aprendido anteriormente deve ser esquecido. Por exemplo, se alguém tem que reaprender a utilizar uma máquina, tem-se a ideia de que a maneira antiga de utilização não será mais empregada. Porém, a ideia que o experimento passa é de uma aprendizagem que ainda não terminou quando o robô aprende que um dos lados do ambiente contém

comida. O aprendizado dele continua quando aprende que o lado da comida pode se inverter.

3.4.4 Crítica

Em geral, nos trabalhos da IA Corporificada, o que se nota é uma desestruturação da área. A ideia de Pfeifer, com seus princípios, é realmente necessária à área. Os conceitos utilizados são ou mal definidos ou sem uniformidade quando se analisa trabalhos de autores diferentes. Ainda há uma carência de discussão teórica na IA Corporificada.

Em particular, é possível notar que grande parte dos pesquisadores da área trata a realidade como objetiva. Ou seja, eles montam seus experimentos sem levar em conta que os diferentes corpos de seres inteligentes vão levar a diferentes compreensões acerca do mundo. Isso ocorre também nos trabalhos que mostramos. No primeiro e segundo experimentos que discutimos, os autores esperam utilizar agente com corpos completamente distintos dos nossos, esperando obterem resultados similares. Como no caso da inversão de sensores ou no comportamento perseverante. No terceiro experimento, conceitos de alto nível, formulados a partir da nossa percepção de mundo são utilizados para explicar o comportamento de um robô que é bem mais simples que nós.

3.5 Críticas à IA Corporificada

As críticas apresentadas aqui foram feitas principalmente por três pesquisadores, Dreyfus (DREYFUS, 2007), Froese e Ziemke (FROESE; ZIEMKE, 2009). Essas críticas têm embasamentos distintos, mas se assemelham. Enquanto Dreyfus continua com sua crítica baseada nos filósofos Heidegger e Merleau-Ponty, afirmando que a IA Corporificada ainda não é "heideggeriana" o suficiente, Froese e Ziemke a partir do paradigma enativista, que abordaremos no próximo capítulo, constroem suas críticas.

Em 2007 Dreyfus publicou um artigo(DREYFUS, 2007) atualizando suas críticas à IA. Assumindo a posição de que as suas críticas à IA Simbólica ainda são válidas, ele apresenta sua visão em relação ao novo paradigma da IA que surgiu com Brooks. Ele chama essa nova IA de "IA Heideggeriana", termo que ele atribui a Wheeler(WHEELER, 2005).

Dreyfus elogia o trabalho pioneiro de Brooks, afirmando que se trata de um avanço para a IA. Porém, ele chama atenção para o fato de que os robôs de Brooks só respondem a características fixas do mundo, eles não aprendem e nem estão inseridos em um contexto. A IA que Dreyfus almeja é uma que resolva o *frame problem*, e, segundo ele, Brooks evita o problema ao deixar de fora significado e aprendizado. Para Dreyfus, o aprendizado não se dá como uma ampliação das possibilidades de ação que a mente tem a seu dispor em determinada situação. Ao invés disso, o corpo é quem aprende a reagir a uma nova situação, aprender é ter a capacidade de deixar o corpo em prontidão para agir diante de uma situação. Por exemplo, uma pessoa que está aprendendo a jogar tênis não o faz de modo a inserir em sua mente novas possibilidades de ação caso a bola se aproxime de determinado modo. Ninguém age a situações desse tipo através de decisões mentais, o corpo deve estar em prontidão para reagir, o corpo aprende os modos de

ação conforme o contexto de como a bola se aproxima. Dreyfus ainda critica Brooks por ter proposto uma produção de inteligência de forma incremental, e ao invés de ter perseguido sua meta inicial da inteligência de um inseto, Brooks deu um salto em sua meta, iniciando projeto para produzir inteligência humana, o projeto Cog(BROOKS et al., 1999).

Dreyfus também critica a proposta de Wheeler (WHEELER, 2005) para a IA. Wheeler, bastante influenciado por Dreyfus, acha que a IA realmente deve se basear em Heidegger, mas que deve ter representações orientadas a ação e a resolução de problemas. Dreyfus diz que reintroduzir a representação significa um retrocesso para um ponto anterior a Brooks. Mas, para um ser com inteligência humana, ele acredita que é necessário que esse ser tenha um corpo muito parecido com o corpo de um humano, e que tenha motivações como as de um ser humano. Para ele, a inteligência depende de fatores corporais e sociais. Então para desenvolver uma inteligência humana, esses aspectos têm que ser satisfeitos como os do ser humano. Mesmo que fosse possível produzir inteligência em um computador de mesa, como ele é hoje, é provável que ele tivesse entendimentos a respeito do mundo diferente dos nossos. O que é relevante para nós poderia não ser relevante para o computador. Ele poderia desenvolver um vocabulário diferente, ou pelo menos um entendimento da linguagem diferente do nosso. A linguagem é uma construção social, o desenvolvimento de linguagem por parte do computador se daria a partir de interações com outros seres. Talvez a real diferença entre a linguagem do computador e a nossa só fosse perceptível a partir da interação de dois computadores similares, que conseguissem perceber como relevantes os mesmos aspectos do mundo. Essa distinção, entre a linguagem humana e uma hipotética linguagem do computador, poderia ser de tal maneira que nos impedisse de perceber o computador como um ser inteligente, mesmo que ele o fosse.

Em 2009 Froese e Ziemke publicaram um artigo(FROESE; ZIEMKE, 2009) que sintetiza a trajetória da IA até chegar à IA Enativa, da qual trataremos no Capítulo 4. Nesse trabalho, eles também fazem críticas importante a respeito das limitações da IA Corporificada.

Froese e Ziemke também criticam os resultados de Brooks, afirmando que seus robôs não resolvem o *symbol grounding problem*, por terem laços sensório-motores rígidos. Froese e Ziemke chamam atenção para o fato de que dizer, como parte dos trabalhos de IA Corporificada fazem, que um sistema é orientado a um objetivo (*goal-directed*) pode ir de encontro ao princípio P5, que fala sobre a questão das perspectivas de análise do agente, já que esse objetivo pode ser aparente para quem olha, mas não intrínseco ao sistema. Por exemplo, um termostato é orientado ao objetivo de controlar a temperatura, mas somente de um ponto de vista externo ao termostato, o objetivo de que o termostato controle a temperatura é de quem o utiliza. Pois, para o próprio termostato, não existe objetivo, ele é um aglomerado de peça compostas de tal forma que a temperatura é controlada, mas nenhuma das peças tem o objetivo de funcionar de determinada maneira, nem o termostato com um todo o tem. Afirmações do tipo "o objetivo de X é Y" podem ser errôneas se a propriedade Y em questão tiver sido imposta externamente (*"externally imposed"*), por quem fabricou X, ao invés de ser internamente gerado (*"internally generated"*) pelo próprio X. A ideia de um objetivo internamente gerado será aprofundada no Capítulo 4.

Os seres inteligentes de que temos conhecimento, e seres vivos em geral, possuem o objetivo intrínseco de sobreviver. Então, Froese e Ziemke questionam o que falta a um sistema

feito de laços sensório-motores, para podermos afirmar que ele gera seus próprios objetivos. A resposta para tal questionamento está no modo de ser dos sistemas de laços sensório-motores e dos seres vivos.

Os seres vivos estão constantemente "funcionando" para manter sua própria existência, suas partes funcionam de modo a manter o todo. Eles estão em constante ato de realização de si mesmos, o seu modo de ser é classificado como "*being by doing*[3]". Já os robôs da IA Corporificada têm seus objetivos traçados por um agente externo e, em geral, não têm a necessidade de se preocupar com aspectos ligados à continuidade da sua existência. Ou seja, eles devem cumprir o objetivo para o qual foram projetados, a energia necessária para tal, ou a manutenção de parte defeituosas são preocupações de outro ser. Esses robôs continuam a existir(ser), mesmo que não façam absolutamente nada, ou mesmo até que estejam desligados. Por isso, seu modo de ser é classificados como "*being by being*[4]".

Esse tipo de distinção nos impede de afirmar que um sistema puramente responsivo com um laço de retroalimentação, como é caso dos sistemas da IA Corporificada, têm a capacidade de gerar seus próprios objetivos. No Capítulo 4, abordaremos mais a fundo a importância dessa capacidade para se afirmar que um sistema é inteligente.

[3] Ser através do fazer
[4] Ser por ser

4 INTELIGÊNCIA ARTIFICIAL ENATIVA

No capítulo anterior apresentamos as bases teóricas e críticas à Inteligência Artificial Corporificada. Diante das críticas, percebemos que é necessário voltarmos um passo e olhar de forma ampla para enxergar o caminho que estamos tomando. Com a percepção de que os significados das, e mesmo as próprias, tarefas e objetivos dos seres artificiais devem ser intrínsecos a eles, não podemos impô-las externamente. É fácil aqui perceber um paralelo com seres vivos, os únicos seres dos quais sabemos serem capazes de gerar intrinsecamente seus próprios propósitos.

Nesse capítulo iremos, na Seção 4.1, apresentar a teoria biológica da autopoiese, de Maturana e Varela. Mostraremos como um ser autopoiético é capaz de gerar sua própria identidade autonomamente e também uma normatividade. Na Seção 4.2, falaremos a respeito da teoria de Varela sobre como o ser autopoiético é capaz de constituir uma perspectiva, se utilizando da normatividade gerada pela autopoiese. Ainda na teoria de Varela, mostraremos como, a partir da perspectiva, ele afirma que o ser vivo é capaz de dar significado às suas interações com o meio.

Na Seção 4.3, falaremos da objeção feita por Di Paolo a respeito da teoria de Varela. Para Di Paolo autopoiese não é suficiente para que o organismo dê significado às suas interações com o mundo, ele afirma que a adaptatividade é essencial para tal, nessa seção mostraremos as consequências da adaptatividade.

Esses conceitos relativos à cognição, derivados da teoria da autopoiese, foram incorporados pela Ciência Cognitiva, resultando na Ciência Cognitiva Enativa. Os pesquisadores de IA perceberam que os conceitos da Ciência Cognitiva Enativa se encaixava com a resposta para as críticas da IA Corporificada. Eles trouxeram esse conceitos para a IA, dando origem à Inteligência Artificial Enativa, da qual falaremos na Seção 4.4. Ainda nessa seção, falaremos da proposta de Froese e Ziemke de princípios para a IA Enativa.

4.1 Autopoiese: A forma de organização dos seres vivos

Podemos observar na biologia que, tradicionalmente, a vida é caracterizada de duas formas. Em uma delas, a vida é caracterizada em termos de mecanismos, sem os quais a vida não seria possível, ou pelo menos que está presente em todos os seres vivos conhecidos hoje: o DNA. Na outra visão, o que é essencial à vida são propriedades de alto nível, então só é dito vivo aquilo que nasce, cresce, reproduz-se e morre. Maturana e Varela, na década de 1970, propuseram uma caracterização alternativa às tradicionais da biologia. Sua caracterização da vida é feita em termos da organização sistêmica dos seres vivos. Seu objetivo, entre outros, era identificar, como essencial a vida, uma propriedade que não fosse contingente e fosse capaz de abranger potenciais formas de vida desconhecidas pelo ser humano. Para Maturana e Varela, o que caracteriza um ser vivo é o fato de ele estar em constante processo de autoprodução.

A respeito da organização do ser vivo, Varela diz:

> At present there is no formulation of this organization, mainly because the great developments of molecular, genetic and evolutionary notions in contemporary biology have led to the overemphasis of isolated components, e.g. to consider reproduction as a necessary feature of the living organization and, hence, not to ask about the organization which makes a living system a whole, autonomous unity that is alive regardless of whether it reproduces or not.[1](VARELA, 1991)

A teoria de Maturana e Varela se aplica diretamente às formas mínimas de vida. Considere uma bactéria, ela é composta de organelas, uma membrana e o núcleo celular. Essas estruturas, através de um fluxo de moléculas e energia, produzem componentes que constituem as próprias estruturas. Ou seja, se olharmos para a bactéria como um todo, ela está constantemente se autoproduzindo. Assim, podemos perceber processos de transformação material que se concatenam, e resultam na produção de componentes que realizam esses mesmos processos. A teoria de Maturana e Varela é chamada de autopoiese, *autopoiesis* em grego significa autoprodução. Varela define um sistema autopoiético:

> An autopoietic system is organized (defined as unity) as a network of processes of production (synthesis and destruction) of components such that these components:
> (i) continuously regenerate and realize the network that produces them, and
> (ii) constitute the system as a distinguishable unity in the domain in which they exist.[2](VARELA, 1991)

A autopoiese é um caso particular de auto-organização. Auto-organização é um tipo de fenômeno que dá origem, de maneira espontânea, a um sistema que se mantém enquanto condições específicas são satisfeitas. Os sistemas auto-organizados se caracterizam por existirem longe do equilíbrio. Isto é, eles se encontram em constante propensão a se desfazerem. O tornado é um exemplo de sistema auto-organizado, que se forma espontaneamente em condições atmosféricas de temperatura e pressão específicas. Outro exemplo são as células de convecção, que se formam em um corpo fluido de forma a otimizar o fluxo de temperatura, decorrente da diferença de temperatura entre dois extremos do fluido.

Destacamos dois aspectos importantes a respeito dos padrões e sistemas auto-organizados. O primeiro é que para poder se manifestar, o fenômeno da auto-organização depende do estabelecimento de um conjunto de condições externas. Apesar de determinarem a possibilidade de ocorrência do fenômeno, essas condições não determinam a forma específica do fenômeno.

[1]No momento não existe formulação dessa organização, principalmente por que o grande desenvolvimento das noções moleculares, genéticas e evolutivas na biologia contemporânea levou a uma ênfase exacerbada dos componentes isolados, *e.g.* considerar a reprodução como um aspecto necessário da organização viva e, portanto, não perguntar sobre a organização que faz um sistema vivo um todo, uma unidade autônoma, que está vivo independente de se reproduzir ou não.

[2]Um sistema autopoiético é organizado(definido como unidade) como uma rede de processos de produção (síntese e destruição) de componentes tais que esses componentes:
(i) continuamente regeneram e efetuam a rede que os produz, e
(ii) constituem o sistema como uma unidade distinguível no domínio no qual eles existem.

Essa forma é determinada pelas características e relações internas dos componentes. É isso que motiva o termo "auto-organização". Por exemplo, no caso das células de convecção, o que determina seu formato são as propriedades do fluido, como o estado físico em que se encontra e sua viscosidade, por exemplo. O segundo aspecto é que, tipicamente, quando as condições externas que permitem a manifestação do fenômeno de auto-organização desaparecem, o padrão ou sistema auto-organizado se desfaz. Por exemplo, quando as condições atmosféricas mudam, o tornado se desfaz, ou quando a diferença entre as temperaturas nos lados opostos do fluido aumenta ou diminui muito, as células de convecção cessam.

Autopoiese também é um exemplo de auto-organização. Porém, o sistema autopoiético possui uma característica distinta dos demais sistemas auto-organizados. Di Paolo fala a respeito da diferença entre autopoiese e as demais formas de auto-organização:

> An essential difference between autopoiesis and the rest of the wider class of self-organization is that what is by definition a process of material self-production must as a result generate a self-distinguishing concrete unity and not simply a physical pattern. The unity is self-distinguishing because it is constructed and sustained by its own activity in spite of the equalizing physical tendencies.[3](DI PAOLO, 2005)

Uma vez constituído, o sistema autopoiético resiste à se desfazer. No caso do tornado não existe nada intrínseco a ele que resista à sua extinção, ou que, possivelmente, estimule a continuidade das condições atmosféricas que o geraram, de forma que o próprio tornado não se desfaça. Já uma bactéria sem contato com seus nutrientes não se desfaz de imediato, somente com a falta prolongada destes é que suas partes começam a se desfazer, além de ser capaz de se deslocar em busca de sua fonte de nutrientes. Note que a resistência a ser desfeito nos dá a ideia que o sistema autopoiético possui objetivo intrínseco de continuidade e ainda uma capacidade de ação para alcançar esse objetivo. Nas Seções 4.2 e 4.3 aprofundaremos esses tópicos.

Um organismo autopoiético está constantemente se autoproduzindo: todos os processos internos do sistema funcionam de tal forma que o sistema se mantém. Por exemplo, em uma célula, todas as suas partes atuam para produzir a própria célula. Na autopoiese, cada parte da célula funciona de modo a produzir alguma outra parte da célula, por transitividade cada parte da célula é essencial na sua própria produção. Como dissemos, sistemas auto-organizado, como a autopoiese, existem longe do equilíbrio. Na autopoiese o funcionamento das partes do sistema garantem a continuidade do sistema, já que a degradação de uma parte é suprimida quando outra parte a produz. Se uma parte A do sistema para seu processo de produção, isto faz com que outra parte B, que era produzida por A, se degrade e não seja recomposta, isso fará com que a parte B também pare seu processo de produção, afetando todas as partes do sistema em cadeia. Eventualmente a parte que produz a parte A também é afetada. Assim, nota-se a presença de uma circularidade na autoprodução do sistema autopoiético.

[3]Uma diferença essencial entre autopoiese e o resto da classe da auto-organização é que o que é, por definição, um processo de autoprodução material deve, como resultado, gerar uma unidade concreta auto-distinguível e não somente um padrão físico. Uma unidade é auto-distinguível porque é construído e mantido sua própria atividade, ao invés de tendências físicas equalizadas.

Um sistema autopoiético está sujeito a um constante fluxo material e energético. Na verdade, sua continuidade como sistema auto-organizado depende desse fluxo. É a partir dele que o sistema consegue obter energia para a contínua produção das parte que o constituem. Fisicamente falando, o organismo autopoiético está em constante modificação dos constituintes de suas partes, isto é, as moléculas que o constituem fazem parte desse fluxo material. Apesar de estar sempre se modificando, há algo no ser autopoiético que permanece constante: sua organização. Varela diz "one way to spotlight the specificity of autopoiesis is to think of it self-referentially as that organization which maintains the very organization itself as an invariant."[4](VARELA, 1991).

Como dissemos, a noção de autopoiese retrata formas mínimas de vida. Mas a ideia de uma organização que se mantém invariante através do fluxo material inerente a seus componentes, pode ser ampliada através da noção de fechamento organizacional[5]. Froese e Ziemke definem fechamento organizacional como:

> We shall say that autonomous systems are organizationally closed. That is, their organization is characterized by processes such that
>
> 1. the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and
>
> 2. they constitute the system as a unity recognizable in the space (domain) in which the processes exist.[6](FROESE; ZIEMKE, 2009)

Isto quer dizer que o sistema organizacionalmente fechado é uma rede recorrente de processos interdependentes. Voltando ao exemplo da célula, podemos olhar para suas partes e perceber que elas não estão diretamente produzindo "a célula", mas produzindo as partes que a constituem. Cada parte da célula produz, ou participa da produção de uma ou mais partes, então podemos dizer que esses processos são interdependentes. Porém, a noção de fechamento organizacional pode ser aplicada também a seres multicelulares. O ser humano é um exemplo, nossos órgão são os processos mencionados na definição, que nos constituem como uma unidade. Note que a própria dependência que existe entre os processos determina quem faz parte da célula, ou seja, o que está dentro e o que está fora da célula. Assim, dizemos que o fechamento organizacional confere ao organismo um caráter de identidade. Essa identidade é independente de um observador externo. Manter e sustentar essa identidade caracteriza o sistema como autônomo. Como suas partes atuam em função da manutenção das próprias partes, podemos afirmar que elas funcionam para a manutenção do sistema como um todo. Note que o funcionamento da totalidade das partes, ou seja, o sistema, funciona também para a manutenção

[4]um maneira de destacar a especificidade da autopoiese é pensar sobre ela de modo autorreferencial como a organização que mantém a própria organização invariante.

[5]Froese e Ziemke em (FROESE; ZIEMKE, 2009) afirmam que também é usado hoje o termo fechamento operacional para denotar o fechamento organizacional.

[6]Nós dizemos que sistemas autônomos são organizacionalmente fechados. Isso é, sua organização é caracterizada por processos tais que

1. os processos são relacionados como uma rede, de modo que eles dependam recursivamente um do outro na geração e realização dos próprios processos, e

2. eles constituem o sistema como uma unidade reconhecível no espaço (domínio) no qual os processos exitem.

do sistema. Assim, podemos identificar um propósito intrínseco à existência do sistema, que é o de se manter funcionando, ou se manter vivo. Desse modo, temos que um organismo com a propriedade de fechamento organizacional gera sua própria lei, o que fundamentalmente o caracteriza como autônomo.

A investigação a respeito de seres vivos, e como se dá o surgimento de autonomia nos mesmos, se refletem na IA como resposta às a IA Corporificada. Ainda não temos uma ideia de como reproduzir artificialmente essas características(autonomia, identidade e propósito intrínseco), mas a teoria autopoiética nos indica o caminho pelo qual se pode tentar isolá-las para reproduzi-las. Na Seção 4.2 nos aprofundaremos nas consequências da teoria autopoiética e na discussão a respeito de como um organismo autopoiético consegue dar significado às suas interações com seu meio.

4.2 Biologia da intencionalidade

Até agora tratamos dos sistemas autopoiéticos como uma unidade em si, mas não falamos da sua relação com seu meio. Como dissemos na Seção 4.1, em geral, as entidades auto-organizadas dependem que condições específicas dos seus meios sejam satisfeitas para surgirem e persistirem. Porém, com os sistemas autopoiéticos há uma particularidade, eles são capazes de persistirem mesmo que tenham sido desfeitas as condições que lhe deram origem. Para isso, estão constantemente se autoproduzindo, através de um fluxo material e energético. Esse fluxo é compartilhado com seu meio, com o qual o sistema troca moléculas. O processo de autoprodução também confere ao sistema a capacidade de se estabelecer como uma entidade separada de seu meio, isto é, de ter uma identidade. Varela diz sobre a relação do sistema autopoiético com seu meio:

> It is ex-hypothesis evident that an autopoietic system depends on its physico-chemical mileu for its conservation as a separate entity, otherwise it would dissolve back into it. Whence the intriguing paradoxicality proper to an autonomous identity: the living system must distinguish itself from its environment, while at the same time maintaining its coupling;[7](VARELA, 1991)

Apesar de Varela falar em paradoxo, as duas propriedades aparentemente antagônicas são características que se complementam na estrutura autopoiética. O acoplamento com o meio é o que permite a continuidade da entidade e, portanto, sua distinção do meio. Por outro lado, sem a distinção a entidade seria igual ao seu meio e não teria como haver o acoplamento. Note que essa relação de acoplamento do sistema autopoiético com seu meio é assimétrica. Em outras palavras, há uma diferenciação dos papeis das duas entidades que se acoplam, sendo o sistema autopoiético que cumpre o papel ativo no acoplamento. Como dissemos na Seção 4.1, ao se definir como uma unidade, o ser autopoiético simultaneamente define o que fica do lado

[7]É evidente por hipótese que um sistema autopoiético depende de seu meio físico-químico para sua conservação como entidade separada, do contrário ele se dissolveria. Daí a paradoxalidade intrigante própria de uma identidade autônoma: o sistema vivo deve distinguir-se de seu meio, enquanto ao mesmo tempo mantém seu acoplamento;

de fora. Essa divisão só faz sentido do lado de dentro da entidade. A partir dessa divisão há o estabelecimento de uma perspectiva, como diz Varela:

> the autopoietic unity creates a perspective from which the exterior is one, which cannot be confused with the physical surroundings as they appear to us as observers, the land of physical and chemical laws simpliciter, devoid of such perspectivism.[8](VARELA, 1991)

Isso quer dizer que, uma vez estabelecida a unidade autopoiética, seu meio não é, do ponto de vista da própria unidade, a coleção de objetos físico-químicos que existiria sem a presença da unidade. Por exemplo, considere uma bactéria se deslocando em um gradiente de sacarose, ela se desloca em direção a regiões com maiores concentrações de açúcar. O gradiente só é interessante para nós porque a bactéria aponta para ele como relevante. Agora podemos fazer uma distinção entre ambiente e mundo. Ambiente é como o meio se apresenta a um observador sem a referência de uma unidade autônoma. Mundo é como o ambiente se apresenta para o sistema autopoiético. O mundo do organismo autopoiético surge no momento que surge sua identidade, ambos são consequência da dinâmica que rege a autopoiese.

A diferença entre o ambiente e o mundo é o que Varela chama de "*surplus of signification*"[9](VARELA, 1991). De acordo com o biólogo, esse *surplus* infesta o entendimento do ser vivo e da cognição. Ele diz ainda que o *surplus* é a mãe da intencionalidade, no sentido de que ele quem dá condições de o sistema produzir suas intenções. Essa diferença entre ambiente e mundo fica clara ao observamos que, no caso da bactéria citado acima, o açúcar só ganha o significado de comida quando a bactéria está presente e o utiliza para permitir a continuação de sua identidade. Ou seja, o próprio ser autopoiético é quem contextualiza o seu meio através da sua intenção ao interagir com ele.

O mundo do sistema é construído a partir das regularidades das leis que governam o ambiente. Por exemplo a capacidade do açúcar de criar um gradiente e de atravessar a membrana é o que possibilita a bactéria usá-lo para se locomover na direção de maior concentração. As propriedades do ambiente são o que garantem a continuidade do acoplamento. A unidade autopoiética está incessantemente se confrontando com encontros com ambiente, sejam eles perturbações, choques ou o acoplamento, ela os encara de uma perspectiva que não é intrínseca aos encontros. Na verdade, essa perspectiva é dependente do organismo autopoiético. Como diz Varela:

> what is meaningful for an organism is precisely given by its constitution as a distributed process, with an indissociable link between local processes where an interaction occurs, and the coordinated entity which is the autopoietic unity giving rise

[8]a unidade autopoiética cria uma perspectiva a partir da qual o exterior é um, que não deve ser confundido com o meio físico como ele aparece para nós, como observadores, a terra de leis físicas e químicas simplesmente, desprovido de tal perspectivismo.

[9]excedente de significação.

to the handling of its environment[10](VARELA, 1991)

De acordo com Varela, a entidade autopoiética está constantemente gerando significação, que para ele é uma carência do ser vivo. Assim, o significado não é preexistente e a relevância dos encontros com o ambiente tem que ser fornecido "*ex-nihilo*"[11](VARELA, 1991). Ainda de acordo com Varela, a construção do mundo por parte do agente autopoiético é feita sempre a partir de desarranjos (*breakdowns*) na autopoiese. Esses desarranjos podem ser desde mudanças na concentração de algum metabólito, até a ruptura de sua membrana.

4.3 Adaptatividade

Até aqui, nós mostramos como a teoria da autopoiese explica a capacidade de seres vivos de constituírem uma identidade, de se estabelecerem como seres autônomos e de possuírem níveis mínimos de cognição. Para além disso, Weber e Varela(WEBER; VARELA, 2002) afirmam que o organismo autopoiético é dotado de teleologia[12] em dois sentidos. O primeiro sentido de que o funcionamento do sistema autopoiético com a propriedade de fechamento organizacional, garante às partes do sistema um propósito intrínseco de continuidade, como já havíamos mencionado. O segundo sentido de que os desarranjos (*breakdowns*) de autopoiese e consequente esforço de reparação desses desarranjos garantem ao sistema como um todo o propósito de permanência.

Di Paolo(DI PAOLO, 2005) diz que a definição de autopoiese não é clara e sempre necessita de interpretação depois de enunciada. Uma vez que esteja claro como interpretar a definição de autopoiese, ela implica somente no primeiro sentido de teleologia que apresentamos, teleologia das partes do sistema. Para Di Paolo o outro sentido de teleologia não se aplica aos seres autopoiéticos, mas se aplica aos seres vivos. O que indica uma falta de algum aspecto nos seres autopoiéticos para se conseguir de fato um propósito intrínseco ao sistema.

Para Di Paolo, as interpretações de autopoiese dependem de noções intuitivas, que normalmente leva a duas diferentes noções: conservação e homeostase. Cada uma dessas noções possui diferentes consequências. A noção de conservação em autopoiese é a noção de que um sistema autopoiético constituído só possui dois tipos de interação com seu ambiente: um que conserva a autopoiese e outro que a desfaz instantaneamente. Enquanto a conservação leva a um determinismo rígido, homeostase "connotes the existence of active mechanisms capable of managing and controlling the network of processes that construct the organism."[13](DI PAOLO, 2005). Assim, notamos que a ideia de Varela(VARELA, 1991) é equivalente a noção de homeostase apresentada por Di Paolo, já que a ideia dos desarranjos na autopoiese requerem

[10] o que é significativo para um organismo é precisamente dado por sua constituição como um processo distribuído, com uma ligação indissociável entre processos locais, onde uma interação ocorre, e a entidade coordenada, que é a unidade autopoiética, dando origem à manipulação do ambiente

[11] A partir de coisa alguma.

[12] Teleologia é área da filosofia que estuda os fins ou os propósitos. Aqui, ser dotado de teleologia significa ter um propósito.

[13] Conota a existência de mecanismos ativos capazes de gerir e controlar a rede de processos que constrói o organismo.

mecanismos capazes de revertê-los, afim de continuar a autopoiese.

Apesar de a interpretação de Varela(VARELA, 1991) com seu *surplus* de significação levar a ideia de homeostase, Di Paolo afirma que, pela definição, autopoiese é um conceito rígido e só pode ocasionar conservação. Conservação só permite uma normatividade do tipo tudo ou nada. Di Paolo diz que na visão conservativa "There is no room for concepts such as lacks, minor or major breakdowns in autopoiesis: either organization is conserved or it isn't."[14](DI PAOLO, 2005). Di Paolo conclui que deveria haver outro conceito que permitisse noções de gradação. Por exemplo, o caso da bactéria que se move em um gradiente de açúcar, em direção às áreas e maior concentração. Autopoiese não oferece uma explicação do porque a bactéria faz esse deslocamento. Como uma noção conservativa, se a bactéria for autopoiética somente, ela deveria ficar parada, se alimentando enquanto chegasse açúcar até ela, mas nada na definição de autopoiese leva a crer que ela se deslocaria em qualquer direção.

Para que o ser vivo possa lidar com desarranjos da autopoiese, como Varela(VARELA, 1991) afirma, ele deve ter alguma outra característica. Di Paolo(DI PAOLO, 2005) usa o termo *sense-making*[15] para designar o nível mínimo de cognição. Segundo ele, para ir de autopoiese a *sense-making* "it is convenient to put in different terms the contrast between what autopoiesis implies and what sense-making requires."[16](DI PAOLO, 2005). Para ele, o que *sense-making* requer e que autopoiese não implica é adaptatividade, que ele define como:

> A system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,
> 1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,
> 2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity. [17] (DI PAOLO, 2005)

Essa característica permite-nos examinar a noção de autopoiese que fala de homeostase. Porém, é importante destacar que a adaptatividade não decorre da definição de autopoiese, nem se sabe o que dá origem a ela nos seres vivos. A adaptatividade permite ao ser vivo se reconstituir depois de um encontro que ameace sua continuidade e até mesmo evitá-lo no futuro. Ela dá a entidade viva a capacidade de auto-monitoração e de autorregulação.

[14]Não há espaço para conceitos tais como falta, desarranjos na autopoiese pequenos ou maiores.

[15]Gerar sentido.

[16]É conveniente colocar em termos diferentes o contraste entre o que autopoiese implica e o que *sense-making* requer.

[17]Uma capacidade do sistema, em alguma circunstância, de regular seus estados e suas relação com o ambiente com o resultado que, se os estados estiverem suficientemente perto dos limites de viabilidade,
1. Tendências são distinguidas e seguidas dependendo de em quais estados vão se aproximar ou afastar do limite e, como consequência,
2. Tendências do primeiro tipo são aproximadas ou transformadas em tendências do segundo tipo de forma que estados futuros não atinjam o limite com uma velocidade extrema.

Para Di Paolo, um sistema precisa ser autopoiético e adaptativo para poder ter *sense-making*. A autopoiese confere identidade e uma normatividade. Já a adaptatividade permite o sistema apreciar seus encontro com respeito a essa norma de uma forma gradual. Assim, ele pode julgar um encontro como mais ou menos disruptivo, por exemplo. Dessa forma o sistema é capaz de construir significado para as entidades presentes no seu mundo.

4.4 Inteligência Artificial Enativa

A Inteligência Artificial Enativa surge como uma proposta para superar as dificuldades preexistentes na IA. As características dos seres autônomos e adaptativos, discutidas nas seções anteriores, demonstram um possível caminho para a IA. Temos que os seres vivos conseguem sobrepor os problemas da IA. Apesar de não existir nenhuma novidade nessa afirmação, o paradigma da enação nos fornece uma explicação para as característica dos seres vivos que conferem a eles essa capacidade.

Seres autopoiéticos e adaptativos são capazes de construir significado para os elementos do seu mundo gerando o contexto para eles. Isso quer dizer que problemas como o *frame problem* e o *symbol grounding problem* não se apresentam para os seres vivos. O primeiro porque pela própria natureza da autopoiese, só faz parte do mundo do sistema autopoiético aqueles elementos que de alguma forma são relevantes para ele, estes ganham significado sempre dentro de um contexto. Isso resolve também o segundo problema mencionado. Além disso, a anterioridade que constatamos na IA Corporificada não existiria em um possível agente dotado de autonomia e adaptatividade. Não há como o projetista desse agente antever um objetivo, já que objetivo e agente surgem ao mesmo tempo.

Seguindo a ideia de Pfeifer de estabelecer princípios para a IA Corporificada, Froese e Ziemke(FROESE; ZIEMKE, 2009) sugeriram dois princípios para a IA Enativa. Esses princípios, de acordo com os autores, não devem ser vistos como substitutivos dos princípios de Pfeifer(PFEIFER; IIDA; BONGARD, 2005), mas complementos no sentido de se perseguir a meta de implementação de inteligência genuína em um ser artificial. Os novos princípios estão relacionados com autonomia e adaptatividade, estabelecendo-as como características essenciais para a implementação de agentes artificiais enativos. Froese e Ziemke(FROESE; ZIEMKE, 2009) definem:

> **EAI-1**: the system must be capable of generating its own systemic identity at some level of description.
>
> **EAI-2**: the system must have the capacity to actively regulate its ongoing sensorimotor interaction in relation to a viability constraint.[18](FROESE; ZIEMKE, 2009)

EAI-1 tem uma relação direta com a autonomia, ou autopoiese, já que requer que o sistema gere sua própria identidade. Como vimos, é através do funcionamento característico

[18]EAI-1: o sistema deve ser capaz de gerar sua própria identidade sistêmica em algum nível de descrição.
EAI-2: o sistem deve ter a capacidade de regular ativamente suas próprias interações sensório motoras vigentes em relação a uma restrição de viabilidade

de suas partes para produzir o próprios organismo que se estabelece a identidade do sistema. O princípio é propositadamente vago, ao dizer "*at some level of description*". De acordo com os autores, isso se deve ao fato de que não é possível parecer autônomo em todos os níveis de descrição. Por exemplo, no nível de descrição das interações atômicas, não há como se falar de autonomia dos seres vivos, já que todos os objetos da descrição, isto é, os átomos e suas sub-partículas, estão sujeitos às rígidas leis da Física.

EAI-2 está relacionado com a adaptatividade. Nesse princípio também há a presença de um termo vago, no caso a definição de *viability constraint*. Froese e Ziemke afirmam que isso garante a possibilidade de se estudar dois problemas, dependendo da definição do termo. Se essa restrição for imposta pelo construtor do agente artificial, é possível se estudar a dinâmica da adaptatividade de forma isolada e também fenômenos dinâmicos como homeostase e ultraestabilidade. Porém, se a restrição tiver de surgir de modo intrínseco a partir do princípio EAI-1, os autores afirmam que isso constitui o "*hard problem*" da IA Enativa.

Podemos notar, uma vez que explicitamos os princípios que norteiam a IA Enativa, que o trabalho de Di Paolo(DI PAOLO, 2000), mencionado na Seção 3.4, se caracteriza como um trabalho de IA Enativa. Nesse caso, o princípio EAI-2 recai no caso da imposição da restrição, o que possibilita ao autor fazer o estudo da homeostase.

O trabalho de Di Paolo, caracterizado como trabalho de IA Corporificada e de IA Enativa, indica uma sobreposição entre essas duas áreas. O próprio estabelecimento de princípios como complemento dos já existentes indica uma continuidade. A pouca quantidade de trabalhos de IA Enativa não torna mais difícil de se afirmar se essa área compõe verdadeiramente uma nova vertente da IA, ou se a IA Enativa se trata do lado conceitual que faltava a IA Corporificada.

5 CONCLUSÃO E TRABALHOS FUTUROS

O objetivo dessa dissertação foi apresentar as mudanças ocorridas na Inteligência Artificial, através de uma análise dos questionamentos filosóficos que levaram a reformulações no pensamento acerca dos objetivos da IA, analisando as áreas da IA tradicional, IA Corporificada e IA Enativa.

Descrevemos as bases do pensamento a respeito da IA, que remontam da criação de máquinas de calcular mecânicas. Essas máquinas foram os primeiros artefatos que tinham capacidades que, até então, eram exclusivas do intelecto humano. A partir de então, alguns filósofos e escritores começaram a discutir a possibilidade real de um artefato mecânico que se igualasse, intelectualmente falando, a um ser humano. Com o advento do computador, alguns cientistas passaram a utilizá-lo na tentativa de produzir programas que demonstrassem capacidades de inteligência semelhante à de pessoas. Em uma conferência em Dartmouth(MCCARTHY et al., acessado em 2012) entre esses pesquisadores surgiu a área da Inteligência Artificial. A partir do seu estabelecimento, a área cresceu e, eventualmente, os seus pesquisadores foram encontrando fatores limitantes às possibilidades da cognição em computadores. Em paralelo ao desenvolvimento da área, alguns filósofos investigavam as bases conceituais que sustentavam a IA. Dreyfus(DREYFUS, 1975) identificou pressupostos que estavam presentes nos trabalhos da área, na maioria da vezes implicitamente. O filósofo então criticou a possibilidade desses pressupostos serem de fato verdade, concluindo que o projeto da IA, como estava sendo desenvolvido, nunca alcançaria seus objetivos. Outros filósofos como Searle(SEARLE, 1980), Dennett(DENNETT, 1984) e o psicólogo Harnad(HARNAD, 1990) descreveram problemas que são imprescindíveis de um agente resolver para poder ser considerado inteligente.

Apresentamos a área da IA Corporificada, que surgiu como uma espécie de retomada do projeto de implementação de uma inteligência genuína. Com a compreensão de que o corpo tem um papel essencial no processo cognitivo, ele é responsável pela interação do agente com o mundo. A partir desse entendimento, Brooks(BROOKS, 1991) desenvolveu uma arquitetura, visando a construção de robôs, que permitia o desenvolvimento incremental através de camadas. A meta de Brooks era conseguir implementar inteligência comparável à de um inseto, para que depois pudesse desenvolvê-la, até chega à inteligência no nível de seres humanos. Às ideias de Brooks se somaram outras, fazendo a área da IA Corporificada crescer desordenadamente. Pfeifer *et al.*(PFEIFER; IIDA; BONGARD, 2005) propuseram princípios norteadores para a área, com os quais tentam englobar as principais características dos trabalhos da área. Mostramos ainda que a IA Corporificada recebeu críticas de Dreyfus(DREYFUS, 2007) e de Froese e Ziemke(FROESE; ZIEMKE, 2009). Indicando que a inclusão do corpo no processo cognitivo era necessária, mas não suficiente para produção de seres artificiais inteligentes.

Por fim, apresentamos a teoria da autopoiese, que explicava como um ser vivo mínimo é dotado de autonomia e constitui sua própria identidade. Como esses conceitos podem ser entendidos em termos de seres mais complexos, através da definição de fechamento organizacional. Além disso, mostramos como a entidade autônoma constrói significado a respeito do seu meio, através da interação entre os dois. Vimos que Di Paolo(DI PAOLO, 2005) chama a atenção que somente a autopoiese não é suficiente para levar à interpretação que Varela(VARELA,

1991) dá a autopoiese. Para que se possa falar em construção de significado, o agente deve ser, além de autônomo, adaptativo. Só assim ele será capaz de julgar seus encontros com o meio de maneira gradual, dando contexto aos encontros. Por último, discutimos como essas ideias podem ser aplicada na IA para superar as barreiras da IA Corporificada. Assim, mostramos a proposta de Froese e Ziemke(FROESE; ZIEMKE, 2009) para mais dois princípios, complementando o trabalho de Pfeifer(PFEIFER; IIDA; BONGARD, 2005).

5.1 Trabalhos futuros

Pretendemos continuar investigando as possibilidades da IA Enativa, possivelmente nos aprofundando nos aspectos filosóficos da fenomenologia de Merleau-Ponty(MERLEAU-PONTY, 2006), como também no estudo do ser de Heidegger(HEIDEGGER, 1995). Queremos, também, estudar a teoria de Kampis(KAMPIS, 1991) a respeito de sistemas que se auto-modificam, como uma espécie de generalização da auto-organização.

No contexto da IA Enativa, investigamos possíveis desdobramentos da discussão de Di Paolo(DI PAOLO, 2005). Pretendemos analisar a contribuição de um sistema nervoso na interação do agente autônomo com seu meio.

REFERÊNCIAS BIBLIOGRÁFICAS

ASHBY, W. *Design for a Brain: The Origin of Adaptive Behavior*. 2. ed. [S.l.]: Chapman and Hall, 1960.

BATESON, G. *Mind and nature: A necessary unity*. New York: Ballantine Books, 1979.

BROOKS, R. A. Elephants don't play chess. *Robotics and Autonomous Systems*, v. 6, n. 1-2, p. 3–15, 1990.

BROOKS, R. A. Intelligence without representation. *Artificial Intelligence*, v. 47, n. 1-3, p. 139–159, 1991.

BROOKS, R. A. et al. The cog project: Building a humanoid robot. In: *Lecture Notes in Computer Science*. [S.l.]: Springer-Verlag, 1999. p. 52–87.

BUCHANAN, B. G. A (very) brief history of artificial intelligence. *AI Magazine*, v. 26, n. 4, p. 53–60, 2005.

CLARK, A. *Being There: Putting Brain, Body and World Together Again*. [S.l.]: MIT Press, 1998.

CREVIER, D. *AI: The Tumultuous History of the Search for Artificial Intelligence*. [S.l.]: BasicBooks, 1993.

DENNETT, D. C. Cognitive wheels: The frame problem of ai. In: HOOKWAY, C. (Ed.). *Minds, Machines and Evolution: Philosophical Studies*. Cambridge, UK: University Press, 1984. p. 129–151.

DI PAOLO, E. A. *Homeostatic Adaptation to Inversion of the Visual Field and Other Sensorimotor Disruptions*. 2000.

DI PAOLO, E. A. Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, v. 4, p. 429–452, 2005.

DREYFUS, H. *On the Internet*. 2. ed. [S.l.]: Routledge, 2009.

DREYFUS, H. L. *O que os Computadores não podem fazer: Crítica da Razão Artificial*. [S.l.]: A Casa do Livro Eldorado, 1975.

DREYFUS, H. L. Why heideggerian ai failed and how fixing it would require making it more heideggerian. *Artificial Intelligence*, v. 171, n. 18, p. 1137–1160, 2007.

EVANS, T. G. A heuristic program to solve geometric-analogy problems. In: *Proceedings of the April 21-23, 1964, spring joint computer conference*. [S.l.: s.n.], 1964.

FROESE, T.; ZIEMKE, T. Enactive artificial intelligence: Ivestigating the systemic organization of life and mind. *Artificial Intelligence*, v. 173, n. 3-4, p. 466–500, 2009.

GLENBERG, A. M. What memory is for. *Behavioral and Brain Sciences*, v. 20, n. 1, p. 41–50, 1997.

GLENBERG, A. M.; ROBERTSON, D. A. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, v. 43, p. 379–401, 2000.

GOERTZEL, B. *The Hidden Pattern*. [S.l.]: Brown Walker Press, 2006.

HARNAD, S. The symbol grounding problem. *Physica D*, v. 42, p. 335–346, 1990.

HEIDEGGER, M. *Ser e Tempo*. 5ªedição. ed. Petrópolis, Brasil: Editora Vozes, 1995.

IZQUIERDO, E.; HARVEY, I. The dynamics of associative learning in an evolved situated agent. In: *Proceedings of the 9th European conference on Advances in artificial life*. [S.l.: s.n.], 2007.

KAMPIS, G. *Self-modifying systems in biology and cognitive science: a new framework for dynamics, information, and complexity*. [S.l.]: Pergamon Press, 1991.

LAKOFF, G.; JOHNSON, M. *Metaphors we Live by*. [S.l.]: University of Chicago Press, 1980.

LEGG, S.; HUTTER, M. A collection of definitions of intelligence. In: *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms, volume 157 of Frontiers in Artificial Intelligence and Applications*. [S.l.]: IOS Press, 2007. p. 17–24.

MCCARTHY, J. What is artificial intelligence? `http://www-formal.stanford.edu/jmc/whatisai/whatisai.html` visitado em 05/11/2012. 2007.

MCCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. In: *Machine Intelligence*. [S.l.]: Edinburgh University Press, 1969. p. 463–502.

MCCARTHY, J. et al. A proposal for the dartmouth summer research project on artificial intelligence. `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`. acessado em 2012.

MERLEAU-PONTY, M. *Phenomenology of Perception*. New York: Routedge and Kegan Paul, 2006.

MINSKY, M. *The Society of Mind*. [S.l.]: Simon & Schuster, 1988.

NAKASHIMA, H. Ai as complex information processing. *Minds Mach.*, v. 9, n. 1, p. 57–80, 1999.

NEWELL, A.; SIMON, H. A. Computer science as empirical inquiry: symbols and search. *Commun. ACM*, v. 19, p. 113–126, 1976.

PFEIFER, R.; IIDA, F. Embodied artificial intelligence: Trends and challenges. In: *Embodied Artificial Intelligence*. [S.l.: s.n.], 2003.

PFEIFER, R.; IIDA, F.; BONGARD, J. New robotics: Design principles for intelligent systems. *Artificial Life, January 2005*, v. 11, p. 1–2, 2005.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Pearson Education, 2010.

SEARLE, J. R. Minds, brains, and programs. *Behavioral and Brain Sciences*, v. 3, p. 417–457, 1980.

SHAPIRO, L. The embodied cognition research programme. *Philosophy Compass*, v. 2, p. 338–346, 2007.

SLAGLE, J. R. A heuristic program that solves symbolic integration problems in freshman calculus. *J. ACM*, v. 10, p. 507–520, 1963.

VARELA, F. J. Autopoiesis and a biology of intentionality. In: *Proceedings of a workshop on Autopoiesis and Percetion*. [S.l.: s.n.], 1991. p. 4–14.

VARELA, F. J.; THOMPSON, E. T.; ROSCH, E. *The Embodied Mind: Cognitive Science and Human Experience*. [S.l.]: The MIT Press, 1992.

WEBER, A.; VARELA, F. J. Life after kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, v. 1, p. 97–125, 2002.

WHEELER, M. *Reconstructing the Cognitive World: The Next Step*. [S.l.]: MIT Press, 2005.

WILSON, R. *Boundaries of the Mind: The Individual in the Fragile Sciences : Cognition*. [S.l.]: Cambridge University Press, 2004.

WILSON, R. A.; FOGLIA, L. Embodied cognition. In: ZALTA, E. N. (Ed.). *The Stanford Encyclopedia of Philosophy*. Fall 2011. [S.l.: s.n.], 2011.

WOOD, R.; DI PAOLO, E. New models for old questions: Evolutionary robotics and the “a not b” error. In: *Costa (Eds.) Proceedings of the 9th European Conference on Artificial life*. [S.l.: s.n.], 2007.

YU, V. L. et al. Antimicrobial selection by a computer: a blinded evaluation by infectious disease experts. *Journal of the American Medical Association*, v. 242, p. 1279–1282, 1979.