UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Filosofia e Ciências Humanas
Programa de Pós-Graduação em Filosofia

CARLOS HENRIQUE BARTH

O "FRAME PROBLEM": A SENSIBILIDADE AO CONTEXTO COMO UM DESAFIO PARA TEORIAS REPRESENTACIONAIS DA MENTE

Belo Horizonte
2019

CARLOS HENRIQUE BARTH

O "FRAME PROBLEM": A SENSIBILIDADE AO CONTEXTO COMO UM DESAFIO PARA TEORIAS REPRESENTACIONAIS DA MENTE

Dissertação apresentada ao Programa de Pós-Graduação em Filosofia da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Filosofia.

Linha de pesquisa: Lógica, Ciência, Mente e Linguagem

Orientador: Prof. Dr. Ernesto Perini Frizzera da Mota Santos

Belo Horizonte
2019

100
B284f
2019

Barth, Carlos Henrique
O "frame problem" [manuscrito] : a sensibilidade ao contexto como um desafio para teorias representacionais da mente / Carlos Henrique Barth. - 2019.
178 f.
Orientador: Ernesto Perini Frizzera da Mota Santos.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Faculdade de Filosofia e Ciências Humanas.
Inclui bibliografia

1.Filosofia – Teses. 2.Ciência cognitiva – Filosofia - Teses. 3.Inteligência artificial - Teses. 4.Representção mental - Teses. I. Perini-Santos, E. (Ernesto) . II. Universidade Federal de Minas Gerais. Faculdade de Filosofia e Ciências Humanas. III. Título.

FOLHA DE APROVAÇÃO

O "Frame Problem": a sensibilidade ao contexto como um desafio para teorias representacionais da mente

CARLOS HENRIQUE BARTH

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em FILOSOFIA, como requisito para obtenção do grau de Mestre em FILOSOFIA, área de concentração FILOSOFIA, linha de pesquisa Lógica, Ciência, Mente e Linguagem.

Aprovada em 11 de fevereiro de 2019, pela banca constituída pelos membros:

Prof. Ernesto Perini Trizzera da Mota Santos - Orientador
UFMG

Prof. André Abath
UFMG

Prof. Marco Aurelio Sousa Alves
Universidade Federal de São João Del Rei

Belo Horizonte, 11 de fevereiro de 2019.

Ao meu pai, que adoraria ter visto este trabalho concluído.

Agradecimentos

Há várias pessoas sem as quais eu não teria sido capaz de realizar este trabalho. Agradeço à Rochelle, meu amor, sem a qual eu não poderia sequer sonhar em iniciar e levar adiante esse projeto, pelo amor e por acreditar. Ao meu pai, infinita fonte de suporte e orientação. A minha mãe e minha irmã pela torcida e pela compreensão. Ao Ernesto Perini, meu orientador, pelo apoio, confiança e liberdade. Aos professores de filosofia da UFMG, que tanto contribuíram para a minha formação, em especial ao André Abath, pelo incentivo e apoio. Aos amigos da filosofia que, de alguma forma, colaboraram nessa empreitada, fazendo da filosofia algo especialmente estimulante e prazeroso, em especial a Daniel Debarry e Verônica de Souza Campos, pelos cafés, debates e horas felizes. Aos amigos de longa data, Jean Francener, Rafaeli Ianegitz, Tiago Scheuer e Mariana Winck, pela constante parceria e presença, mesmo na distância. Ao Anselmão, pelo brilho nos olhos. Agradeço também aos membros do departamento, pelo apoio em todas as questões administrativas, ao CNPQ e, claro, aos russos.

Lista de ilustrações

Sumário

Resumo

A sensibilidade ao contexto é uma das marcas distintivas da inteligência humana. Compreender o modo flexível como o ser humano pensa e age em função de um número potencialmente infinito de circunstâncias, ainda que munido de recursos finitos e limitados, é um desafio central para a filosofia da mente e para a ciência cognitiva, em particular aos que fazem uso de teorias representacionalistas. Nesse trabalho, adotou-se como fio condutor o modo como isso se manifesta no *frame problem*: a dificuldade em explicar como a cognição humana reconhece, de maneira eficiente, o que é ou não relevante em cada contexto. A partir dele, buscou-se caracterizar uma tensão fundamental entre a sensibilidade ao contexto e o uso de representações mentais em teorias da cognição. O primeiro capítulo discute a natureza do *frame problem*, bem como as razões de sua resiliência. No segundo e terceiro capítulos, faz-se uso do problema como métrica para investigar o quão adequado é o tratamento das dependências contextuais no âmbito de várias abordagens representacionais. No decorrer da discussão, realiza-se um esforço argumentativo para mostrar que 1) nenhuma das estratégias abordadas é capaz tratar adequadamente da sensibilidade ao contexto, mas que 2) apesar disso, o *frame problem* não constitui argumento fatal para teorias representacionalistas em geral, e que 3) ele constitui uma ferramenta conceitual fundamental para pesquisas contemporâneas.

Palavras-chave: frame problem. contexto. racionalidade. veículo do pensamento. inteligência artificial. representação mental. ciência cognitiva.

Abstract

Context sensitivity is one of the distinctive marks of human intelligence. Understanding the flexible way in which humans think and act in a potentially infinite number of circumstances, even though they're only finite and limited beings, is a central challenge for the philosophy of mind and cognitive science, particularly in the case of those using representational theories. In this work, the *frame problem*, that is, the challenge of explaining how human cognition efficiently acknowledges what is relevant from what is not in each context, has been adopted as a guide. By using it, we've been able to describe a fundamental tension between context sensitivity and the mental representations used in cognition theories. The first chapter discusses the nature of the frame problem, as well as the reasons for its persistence. In the second and third chapters, the problem is used as a measure tool in order to inquiry a few representational approaches and check how well suited they are to deal with context dependencies. The problems found are then correlated with the frame problem. Throughout the discussion, we try to show that 1) none of the evaluated approaches is capable of dealing with context sensitivity in a proper manner, but 2) that's not a reason to think that the frame problem constitutes an argument against representational approaches in general, and 3) that it constitutes a fundamental conceptual tool in contemporary research.

Keywords: frame problem. context. rationality. vehicle of thought. artificial intelligence. mental representation. cognitive science.

Introdução

> Nada é mais admirável que a rapidez com que a imaginação sugere suas ideias, apresentando-as no instante em que elas se tornam necessárias ou úteis. A fantasia percorre o universo de um extremo ao outro, reunindo as ideias que dizem respeito a um determinado assunto. É como se a totalidade do mundo intelectual das ideias fosse a um só tempo exposta à nossa visão, e simplesmente escolhêssemos as mais adequadas a nosso propósito. No entanto, as únicas ideias que podem estar presentes são aquelas que foram reunidas por essa espécie de faculdade mágica da alma, a qual, embora seja sempre a mais perfeita possível nos grandes gênios - constituindo, aliás, precisamente o que denominamos gênio - permanece inexplicável para o entendimento humano, a despeito de todos os seus esforços. (HUME, 2000, p. 48)

A campainha tocou. Como descobrir qual a coisa adequada a se fazer diante desse evento? Feita dessa forma, a pergunta faz com que se comece a pensar: que espécie de resposta é esperada? Afinal, é possível responder de vários modos. Em termos gerais, seria possível dizer que, diante de uma campainha que toca, espera-se que alguém verifique quem é que está à porta. Nessa resposta, descreve-se aquilo que as pessoas rotineiramente fazem, de modo parcialmente descolado de quaisquer cenários concretos. Mas é também possível dar uma resposta restrita a circunstâncias mais específicas: a porta pode ser aberta sem checagem prévia, pois dado o pouco tempo decorrido, com certeza a pessoa que está à porta é a mesma pessoa que acabou de ligar avisando que estava prestes a chegar.

A incerteza sobre o tipo de resposta adequada se dá porque, aqui, a pergunta foi feita sem oferecer um cenário prévio no interior do qual ela possa ser analisada. Não por acaso, a resposta a uma pergunta feita nestas condições faz uso frequente de termos como “depende” e de considerações acerca das conjunturas possíveis, no interior das quais se poderia formular respostas adequadas O fenômeno interessante, porém, é precisamente o de que, para qualquer conjunto de circunstâncias oferecido, a resposta adequada soará razoavelmente óbvia. O ser humano é sensível às circunstâncias em que se encontra e, em geral, parece capaz de levá-las em conta de modo rápido e sem que se demande qualquer esforço considerável de sua parte. Essa é uma daquelas capacidades que, de tão triviais, demandam esforço para enxergar como, por que, e para quem elas podem constituir um problema. Não obstante, a tentativa de tornar não misteriosa essa capacidade aparentemente tão banal está por trás de um dos maiores desafios enfrentados por vários filósofos e cientistas no interior da filosofia da mente e da ciência cognitiva.

Por que explicar essa habilidade constitui um desafio? Fundamentalmente, porque para estes filósofos e cientistas, a mente humana é um tipo de computador. Tal crença é bastante difundida hoje, tanto por eles, quanto por parte da comunidade de cientistas da computação. Se ela for verdadeira, então deve ser possível não ape-

nas fazer uso de modelos computacionais para descrever o modo como as pessoas pensam, mas também fazer com que computadores possam pensar tal como seres humanos o fazem. Dada a familiaridade desse tipo de afirmação, nem sempre é fácil mensurar o quão fortes e demandantes são os pressupostos envolvidos. O que torna teses como esta aceitáveis? O motivo tem pouca relação com os feitos de engenharia no âmbito das ciências da computação. O elemento fundamental é antes uma certa concepção de mente cuja origem é rastreável até a modernidade. Nos tempos de Descartes, percebeu-se que a matemática não descreve objetos ou forças físicas de nenhum tipo específico, mas sim relações abstratas entre representações simbólicas.[1] Descrições geométricas ou físicas, por exemplo, seriam apenas aplicações da matemática. Quando Descartes tomou o próprio pensamento como um tipo de representação simbólica, abriu-se o caminho para que as conclusões acerca da matemática servissem também para a própria mente. Na concepção resultante, pensar é manipular símbolos que representam, mas que não estão intrinsecamente ligados à coisa alguma no mundo.

Um longo tempo depois, na contemporaneidade, com o surgimento dos computadores, surgiu também o *cognitivismo*, isto é, a ideia de que modelos computacionais poderiam ser utilizados para descrever processos mentais. A perspectiva cognitivista, somada a uma certa concepção de inteligência, deu origem às pesquisas em Inteligência Artificial na sua versão clássica. Nelas, a inteligência humana era fruto de sua capacidade de pensar, e esta, por sua vez, era explicada de modo fiel à concepção cartesiana: pensar é gerar e manipular representações simbólicas de acordo com certas regras formais.

Se pensar é manipular símbolos, e se o comportamento inteligente é fruto de um certo modo de manipular símbolos, segue-se que um computador deve ser capaz de agir adequadamente diante de uma campainha que toca. Contemporaneamente, há computadores que jogam xadrez, que dirigem carros e que controlam aeroportos de modo igual ou superior ao que um ser humano seria capaz de fazer. Contudo, ainda hoje, não há um computador sequer capaz de atender uma campainha. Ao menos não um que o faça de modo inteligente, flexível e sensível às circunstâncias, tal como um ser humano faria enquanto boceja. O que há são computadores capazes de agir no interior do conjunto de circunstâncias para os quais foram programados, e somente nele.

Foi no âmbito das pesquisas em inteligência artificial que se originou o problema sobre o qual esse trabalho orbitará: o *frame problem*. Ele pode ser sintetizado, ainda que de forma vaga, como o problema de explicar a capacidade que o ser humano tem para fazer intercâmbio entre diferentes contextos de um modo que seja tratável por teorias da mente com caráter computacional. De forma um pouco menos vaga, trata-se do problema de explicar como um sistema computacional é capaz de

[1] Conforme HAUGELAND (1989).

se mostrar sensível àquilo que é *relevante* em uma dada circunstância. Dado o tocar da campainha, é possível inferir que ela está funcional, que ela não foi roubada, que a chuva do dia anterior não foi capaz de afetá-la, que alguém lá fora possui habilidade, altura e peso suficientes para apertá-la, e assim por diante. Diante desse tipo de inferência possível, duas coisas se sobressaem: primeiro, o quão irrelevantes elas são. Segundo, o quão relevantes elas podem ser. Basta que qualquer uma delas seja considerada no interior do cenário adequado para que ela se mostre como relevante. O ser humano parece ter a capacidade de reconhecer os diferentes cenários em que cada uma destas possibilidades se mostra saliente. Contudo, se tal capacidade for exigida de um sistema computacional, o resultado é um processo potencialmente infinito. Isso porque a distinção entre o que é ou não relevante só parece realizável por meio da checagem explícita de toda e qualquer possibilidade. Para evitar a necessidade de uma tal checagem universal, é preciso, portanto, estabelecer algum modo de circunscrever a gama de possibilidades relevantes de acordo com cada cenário sem checá-las.[2]

A pouca clareza de uma formulação como esta leva à necessidade de uma formulação não sintetizada. Este é objeto do primeiro capítulo. Como se verá, além de ser um desafio para teorias computacionais da cognição, o *frame problem* constitui também um desafio para quem tenta descrevê-lo: parece haver uma gangorra entre explicações claras e explicações curtas. Por ora, convém então apenas enfatizar que não se trata de um problema tecnológico. Ele não pode ser resolvido por apelo a formas computacionalmente mais poderosas de fazer o tipo de coisa que computadores atuais já fazem, ou ao menos para esse fim se argumentará. Isso é o que faz dele um problema importante tanto para o cientista quanto para o filósofo: trata-se de uma barreira tanto para quem quer fazer um computador pensar de modo inteligente, tal qual um ser humano, quanto para quem quer "apenas" explicar a inteligência e o pensamento humanos.

No decorrer da investigação, ao mesmo tempo em que se tenta elucidar alguns aspectos mais profundos desse desafio, será possível mostrar como ele se manifesta em diversas teorias computacionais da mente. Ainda no primeiro capítulo, será explorada a tese de que o problema é fruto do uso de sistemas formais para realizar a modelagem dos processos mentais. As discussões orbitarão, portanto, os diferentes resultados obtidos pelo uso de formalismos distintos, da lógica clássica a formalismos pouco usuais, tais como lógicas não monotônicas. As dificuldades enfrentadas por estas primeiras tentativas de solução serão representativas do tipo de dificuldade que se encontrará nas demais abordagens posteriormente tratadas.

A partir do segundo capítulo, já de posse de uma concepção mais detalhada e específica do *frame problem*, será possível começar a fazer uso dele como ferramenta de análise de outras teorias. Como se verá, as dificuldades enfrentadas pelas

[2] Isto é, de colocar um *frame* ao redor do conjunto de elementos relevantes, separando-os dos irrelevantes.

primeiras abordagens levaram filósofos e cientistas a buscar a natureza do problema em outras águas. Para alguns deles, ele não seria fruto do uso de um formalismo ou outro, mas sim do uso de um certo gênero representacional. De acordo com esta hipótese, as representações mentais tipicamente usadas nas teorias computacionais teriam caráter linguístico, e é em virtude desta estrutura que o *frame problem* viria à tona. Duas alternativas se mostram: o uso de representações distribuídas, típicas de abordagens conexionistas, tais como redes neurais; e o uso de representações pictoriais, de natureza imagética. O capítulo dois busca tratar destas possibilidades.

No capítulo três, será explorada a tese de que haveria um pressuposto implícito na concepção de inteligência perseguida pelas teorias computacionais: a racionalidade. Segundo essa hipótese, teorias computacionais que sofrem do *frame problem* estariam fazendo uso de uma concepção de racionalidade excessivamente idealizada, resultando na demanda por criaturas ideais, e não criaturas finitas e imperfeitas como o ser humano. Bastaria, portanto, substituir a concepção de racionalidade adotada por uma menos idealizada e o problema estaria solucionado.

Em termos gerais, será realizado um esforço argumentativo para sugerir que o problema sobrevive a todas essas abordagens: ele permanece sem solução. Em todos os capítulos, será possível observar uma mesma metodologia: primeiro, a apresentação de uma tese que sintetiza o modo como o problema tentará ser resolvido. Segundo, a consideração de tentativas concretas de formular teorias em conformidade com estas teses. Na sequência, serão apresentadas então as sugestões de como e do porquê estas teorias não são satisfatórias. Desse modo, será possível perpassar autores diversos, mas sempre fazendo uso de um único fio condutor, espera-se, bem constituído.

Como o problema permanece sem solução, o valor desse tipo de investigação está, acredita-se, no quão instrutiva ela pode ser, tanto para aqueles que querem resolver o problema, quanto para aqueles que buscam fazer uso dele como argumento contra abordagens computacionais. Ao mesmo tempo, será possível salientar o valor do próprio *frame problem* como uma ferramenta para explorar os limites e desafios inerentes à filosofia da mente e à ciência cognitiva. Num certo sentido, o papel por ele exercido nessa empreitada é análogo ao do ceticismo na epistemologia: talvez seja possível resolvê-lo em definitivo, talvez não, mas de algum modo, ele se faz presente, guiando as investigações.

1 Apresentação do Frame Problem

Como é possível explicar o fenômeno da atividade mental? Esta pergunta pode ser compreendida de pelo menos dois modos distintos. Por um lado, ela pode expressar um desafio localizado no interior de um paradigma explicativo. A física, por exemplo, faz uso de explicações nomológico-dedutivas, expressando regularidades na forma de leis. A empreitada de explicar a mente seria, neste sentido, análoga à do físico que tenta elaborar explicações deste tipo para os fenômenos. Contudo, é possível questionar quão adequadas são as explicações de tipo nomológico-dedutivo para o caso da mente. Isso leva a um modo relacionado, porém distinto, de interpretar a questão: qual o paradigma explicativo adequado para tratar da atividade mental? É possível abarcar todos os fenômenos mentais expressando regularidades em forma de leis, sejam quais forem? De modo geral, a ciência cognitiva nega esta possibilidade. Ela não está sozinha nisso. Segundo DUPRE (2012), filósofos da biologia contemporâneos tem se mostrado céticos em relação à existência de leis biológicas, preferindo falar em modelos biológicos. Por sua vez, a ciência cognitiva faz uso de explicações que não expressam regularidades gerais ou universais, mas sim o modo como operam os mecanismos que subjazem aos fenômenos. Tais mecanismos são, em geral, descritíveis em termos nomológico-causais mas, por si só, tais descrições carecem de poder explicativo. Assim, se interpretada da primeira forma, ela constituirá uma pergunta de âmbito científico, isto é, sobre o mundo. Já no segundo sentido, trata-se de uma pergunta acerca do modo como o mundo deve ser investigado, flertando de modo mais direto com a filosofia da mente e da ciência.

A chamada *revolução cognitiva* da década de 1950 foi, dentre outras coisas, uma tentativa de responder a esta pergunta no segundo sentido. Um dos objetivos do movimento intelectual era defender a validade de um certo tipo de explicação para o mental. De lá para cá, muitos dos pressupostos deste movimento vêm sendo atacados, e há quem defenda a necessidade de uma refundação.[1] Este capítulo tem como objetivo principal fornecer a descrição de um problema que tem potencial para afetar toda a empreitada cognitiva: o *Frame Problem*.[2] A pergunta “o que é o FP?” vem recebendo diferentes respostas desde sua primeira elaboração em 1969. Além de ser um problema de difícil solução, é também um problema de difícil enunciação. Em geral, tentativas de apresentá-lo adequadamente oscilam entre versões acuradas de difícil compreensão e versões acessíveis, mas que dificultam uma apreciação adequada da

[1] Ver, por exemplo, CLARK (1998) e WHEELER (2007).

[2] Deste momento em diante, FP. Optou-se por não traduzir o termo “frame problem” por duas razões: primeiro, termos como “estrutura”, “moldura” ou análogos não representam adequadamente nenhum dos sentidos em que este termo figura quando utilizado na literatura inglesa pertinente. Segundo, porque mesmo na literatura em português o termo costuma ser mantido sem tradução, e a quebra desta prática poderia constituir uma dificuldade na compreensão do texto.

enormidade do desafio que ele apresenta. Não por acaso, no decorrer da década de 1980, mais do que uma possível solução, a elucidação do tipo de problema que o FP é, foi tema de um intenso debate entre pesquisadores de inteligência artificial e filósofos. Neste debate, a quantidade de diferentes concepções era próxima da quantidade de debatedores. Por isso, mesmo hoje, raro é o autor que se arrisca a falar do FP sem antes descrever o modo como o compreende.

Dado esse cenário e o desafio de apresentá-lo de um modo, espera-se, compreensível, mas sem deixar de fazer jus à sua complexidade, optou-se por utilizar sua história como guia: busca-se resgatar o contexto de sua origem enquanto problema lógico para só então discutir os desdobramentos que fizeram dele um problema filosófico. Primeiro, será descrito o contexto geral em que o FP surgiu, perpassando o cognitivismo, o uso de modelos computacionais para explicar a mente, bem como o surgimento da *Inteligência Artificial*[3] em sua versão clássica. A partir daí, será caracterizado o contexto específico em que ele se mostrou como problema técnico e serão apresentadas algumas das principais tentativas de solução. Estas são especialmente importantes porque deixam clara a enorme resistência do problema, permitindo compreender por que muitos autores sentem-se justificados em fazer, a partir dele, afirmações ousadas sobre a natureza e a plausibilidade da abordagem cognitivista.[4] Por fim, será descrito o modo como ele extrapolou este contexto específico, vindo a constituir um problema filosoficamente relevante e os desdobramentos que daí advém.

1.1 Em busca dos mecanismos da mente

O *cognitivismo* é uma doutrina cuja origem está relacionada ao surgimento dos computadores. Ele envolve concepções específicas da mente e inteligência humanas. Assim como o motor de um carro pode ser compreendido como um sistema de conversão de energia, a mente é tomada como um *sistema de processamento de informações*. Trata-se de um sistema capaz de receber informações (*input*), realizar operações com elas e oferecer outras informações na saída (*output*). Abordagens cognitivistas pautam-se, primeiro, por distinções bem definidas entre o que está "dentro" e o que está "fora" do sistema e, segundo, por especificações de como se dá o fluxo de entrada, trabalho (processamento) e saída de informações. A entrada de informações no sistema cognitivo a partir do ambiente, por exemplo, pode ser descrita por meio de subsistemas perceptuais bem definidos. Os *transdutores* de FODOR (1983) são subsistemas deste tipo: o sistema visual, por exemplo, recebe fótons do ambiente e converte-os em representações simbólicas dos objetos presentes. Tais representações são então disponibilizadas em um formato adequado para manipulação interna, figurando em juízos e raciocínios. Um processo similar ocorre no sentido

[3] Daqui por diante, IA.

[4] Para Fodor, por exemplo, o FP é um dos motivos pelos quais não é possível uma ciência que tenha como objeto os processos centrais da mente (1983).

oposto: transdutores são responsáveis por "converter" elementos simbólicos em elementos causais. Evidentemente, uma tal descrição mantém o mistério sobre como esta conversão é possível. Não por acaso, John Haugeland afirmou que a função dos transdutores em sistemas cognitivistas é *". . . a função que Descartes assinalou à glândula pineal"* (1998a, p. 223). Contudo, devido ao caráter sistemático[5] das explicações destes mecanismos, uma descrição detalhada de como cada componente realiza sua tarefa não é fundamental. Para o cognitivista, é suficiente assumir que tal processo seja realizável de algum modo (possivelmente mais de um), visto que todo ser humano é capaz de realizar juízos e agir em função deles, sem deixar de estar sujeito às leis da natureza.

Isso permite explicar a relação entre pensamento e ação, bem como a relação entre um evento (o tocar da campainha) e o comportamento do agente diante do evento (dirigir-se à porta para recepcionar alguém). O evento fornece uma miríade de *inputs* perceptuais. A mente realiza então uma série de operações com estes *inputs*, articulando associações e derivando consequências destas associações. O resultado são informações "processadas" capazes de explicar o comportamento do agente (a crença de que há alguém à porta, por exemplo). Algo análogo ocorre em todos os cenários nos quais a mente exerce algum papel. Assim, para o cognitivista, explicar o modo como a mente realiza o processamento das informações permitirá elucidar a participação da mente nestes cenários.

Em sua empreitada, o cognitivista faz uso de um tipo específico de explicação. Em vez das explicações nomológico-dedutivas, típicas da física, o cognitivismo faz uso do que Haugeland denominou *explicação sistemática* (1998b). Nela, a operação ou habilidade mais geral (processar informações no caso da mente, converter energia de tipos específicos no caso de um motor) é explicada por meio de uma descrição estrutural.[6] Em vez de elaborar leis, o cientista busca descrever os componentes do sistema e o modo como estes se articulam. Cada componente é capaz de realizar alguma operação, mas esta é sempre menos geral, ou seja, mais específica e menos flexível. Disso emerge a necessidade de explicar como são possíveis as operações realizadas por estes componentes. Esses são então tomados como subsistemas que, por sua vez, também serão explicados a partir da especificação dos componentes que os constituem e do modo como se estruturam. O mesmo se dá com os componentes destes subcomponentes, cujas operações são cada vez mais específicas e menos

---

5 Esclarecimentos sobre o que são explicações sistemáticas serão dados em seguida.

6 Explicações sistemáticas são um tipo de explicação por modelos, tais como aqueles que, segundo DUPRE (2012), são largamente utilizados na biologia. Em particular, a caracterização das explicações sistemáticas feita por Haugeland é muito semelhante ao modo como BECHTEL (2008) caracteriza explicações mecanicistas em geral. Contudo, na literatura ligada à Inteligência Artificial, é possível observar afirmações como a de que mecanismos são descrições que podem ser explicadas em termos nomológico-dedutivos. Assim, o uso da expressão "explicação mecanicista" pode causar alguma confusão. A fim de evitá-las, optou-se por adotar o termo "explicações sistemáticas" conforme sugerido por Haugeland.

flexíveis.[7]

Contudo, esta caracterização morfológica não é suficiente para constituir uma explicação sistemática. Uma analogia com ambientes empresariais pode ser útil: tudo aquilo que a organização realiza (não apenas o produto que ela gera, mas todos os meios pelos quais ela interage com elementos externos tais como clientes e fornecedores) não é fruto do simples somatório da atividade exercida por cada departamento, mas sim do modo como estes interagem entre si e de suas interdependências. No mesmo sentido, uma explicação sistemática demanda um elemento adicional: interação cooperativa organizada.[8] Tal interação constitui uma interdependência particular entre os componentes. É desta interação que a habilidade mais geral a ser explicada emerge. Fornecer explicações sistemáticas da mente não se resume, portanto, a identificar e descrever suas partes. É preciso identificar seu papel funcional e o modo como este afeta os demais componentes e subcomponentes. O espaço da explicação sistemática termina quando as operações dos componentes são tão específicas e inflexíveis que podem ser explicadas de outras formas, a depender do veículo físico que constitui o sistema (circuitos elétricos, tecido orgânico etc). Nesse cenário, a mente pode ser tomada como um sistema sem que se abra mão da possibilidade de redução à explicações nomológico-dedutivas.

Caracterizada a explicação sistemática, resta agora tratar do modo como o cognitivismo compreende o comportamento inteligente. Na perspectiva cognitivista, ser inteligente é ser capaz de resolver problemas. Diante de um determinado estado de coisas inicial e um estado de coisas almejado, o agente que demonstra a capacidade de planejar ou elaborar “soluções” que levem de um estado ao outro manifesta inteligência. Como isto se encaixa na concepção cognitivista da mente? Sendo a mente um sistema que processa informações, então a inteligência exibida deve ser fruto deste processamento. A tese cognitivista é a de que o comportamento inteligente pode ser completamente explicado por meio de processos cognitivos, isto é, por meio do exercício das operações dos componentes constitutivos da mente. Em síntese, a mente humana é um sistema capaz de processar informações com base em conjuntos de regras e princípios de aplicação de um modo tal que resulta em comportamento inteligente.

Para algumas linhas de pesquisa contemporâneas, as premissas cognitivistas adotadas pela IA podem surpreender. A crença de que um computador pode apresentar inteligência igual ou análoga à inteligência humana parece envolver a suposição de que o corpo não tem participação alguma na constituição das capacidades men-

[7] Sobre a noção de especificidade em jogo, é importante notar que ela se refere a aspectos do formalismo utilizado para descrever o sistema e não ao modo como as habilidades ou operações são cotidianamente descritas. Embora esses possam coincidir, como quando se descreve a operação de multiplicar como a repetição da operação de somar, este não é sempre o caso. Empurrar um objeto é mais específico que mover um objeto, visto que empurrar é um dos possíveis modos pelos quais se pode mover algo, mas disto não se segue que a habilidade de empurrar seja um subcomponente estrutural da habilidade de mover algo.

[8] Conforme HAUGELAND (1998b).

tais. Contudo, expressa nestes termos, esta não é uma acusação justa. Cognitivistas não negam que o comportamento inteligente do ser humano envolve elementos afetivos, motivações e expectativas. Além disso, eles sabem que a interação contínua entre estes elementos e o ambiente ao redor do indivíduo por meio da percepção e da ação são igualmente importantes. Acusá-los de desdenhar do papel do corpo humano na explicação da inteligência humana é impreciso. O que os caracteriza é um certo modo de conceber a relação entre os mecanismos que realizam as operações mentais responsáveis pelo comportamento inteligente e outros mecanismos, como aqueles ligados aos afetos e à percepção. Perceber o mundo, por exemplo, é concebido como a recepção de informações para processamento. Uma descrição cognitivista da percepção é a descrição de uma *interface*, um elemento necessário, mas funcionalmente distinto. Assim, o cenário cognitivista incentiva a concepção dos mecanismos mentais como constitutivos de um sistema funcionalmente isolado, cujas relações com os demais subsistemas ocorrem por *interfaces* bem definidas. Nesse sentido, não é o desinteresse pelo papel do corpo, mas sim a adoção desta perspectiva sistemática o que fundamenta a expectativa de que é possível introduzir em computadores as operações específicas responsáveis pela inteligência de um agente humano.

Esta compreensão de inteligência permeia quase todas as pesquisas em IA. O marco inicial desta empreitada científica deu-se em 1956, quando um conjunto de pesquisadores se reuniu em Darthmouth para somar esforços em torno de uma certa tese. John McCarthy a descreveu nos seguintes termos: *"(. . .) todos os aspectos da aprendizagem ou qualquer outra característica da inteligência podem, em princípio, ser descritos tão precisamente que uma máquina pode ser feita para simulá-los."* (RUSSELL; NORVIG, 2010, p. 17). A "descrição precisa" que McCarthy tinha em mente era o tipo de descrição que figura em explicações sistemáticas. Assim, nesta conferência foi dada a largada para uma empreitada de caráter cognitivista cujo foco estava na simulação do comportamento inteligente.

Evidentemente, o cognitivismo por si só não poderia viabilizar uma empreitada como a da IA. Fosse ou não verdadeira a tese cognitivista acerca da mente, restaria ainda saber que tipo de máquina é capaz de simular sistemas de processamento de informações. Por isso, para a IA, tão importante quanto o surgimento do cognitivismo foi o advento dos computadores. O que os diferencia de outras máquinas quaisquer? Adota-se aqui a tese de HAUGELAND (1989), segundo quem computadores são um tipo de *sistema formal automático*. Grosso modo, um sistema formal é composto por dois conjuntos de especificações: quantos e quais serão os tipos de símbolos disponíveis, e quantas e quais serão as regras de movimentação (i.e. as operações permitidas) destes símbolos. As regras de um jogo de xadrez, por exemplo, constituem um sistema formal. O xadrez trabalha com um número determinado de símbolos (ou seja, de peças). Estes símbolos podem ser de diferentes tipos (peão, bispo, cavalo etc.). Além disso, existem regras de movimentação específicas para cada tipo (o bispo, por exemplo, só pode se movimentar transversalmente). Estas regras de movimento levam

em conta tanto o tipo do símbolo quanto o atual estado do sistema (isto é, a atual disposição das peças no tabuleiro). De modo análogo, computadores também possuem especificações de tipos de símbolos e regras que regem as operações possíveis sobre estes símbolos. Além disso, possuem estados internos e realizam operações sobre os símbolos (i.e. movimentos de "peças") que levam estes estados em conta. Sendo sistemas formais, tanto o xadrez quanto computadores podem ser fisicamente implementados de diversos modos. É possível jogar xadrez tanto com peças de madeira sobre um tabuleiro quanto por meio de anotações sobre papel e caneta. Do mesmo modo, computadores podem ser implementados em qualquer meio que disponibilize os fenômenos físicos necessários, sejam estes mecânicos, hidráulicos ou elétricos.

Dizer de um sistema formal que ele é automático significa dizer que as operações são realizadas sem qualquer controle externo direto. Contudo, a introdução da automaticidade traz à tona a pergunta: quais serão as jogadas realizadas pelo sistema? Há vários modos de iniciar uma partida de xadrez, assim como há vários modos de responder às jogadas adversárias. Na ausência de um agente, como um sistema formal automático "decide" por conta própria qual será a próxima operação a realizar? A resposta é dada pelo estado atual do sistema. No caso do xadrez, o estado do sistema é dado pela disposição das peças no tabuleiro. O mesmo se dá no caso da mente cognitivista. A forma como alguém reage ao tocar de uma campainha é, grosso modo, fruto do conjunto de disposições internas que ele tem e das circunstâncias em que se encontra.

Quem demonstrou a possibilidade de um sistema formal automático nestes moldes foi Alan Turing (1936) por meio do modelo computacional que ficou conhecido como *máquina de Turing*. A tese de Turing é que, para qualquer sistema formal determinístico, há um modelo de uma máquina de Turing que pode comportar-se de modo equivalente.[9] Segue-se disso a possibilidade de modelar uma máquina de Turing que consiga realizar todas as operações necessárias para jogar xadrez adequadamente. A importância dessa tese para a IA é clara: se a mente humana pode ser sistematicamente explicada e, se há um tipo de máquina capaz de comportar-se de modo equivalente à qualquer sistema formal, então há uma máquina de Turing capaz de apresentar o mesmo tipo de comportamento que a mente humana apresenta, inclusive comportamento inteligente. Esse é, em síntese, o princípio por trás de uma teoria computacional da mente de caráter cognitivista, doravante denominada *teoria computacional da cognição* (TCC).

O uso de modelos computacionais traz consigo uma consequência importante para a empreitada da IA: a irrelevância do significado dos símbolos na determinação dos estados do sistema formal. Um outro modo de dizê-lo é que as operações consideram somente propriedades formais ou sintáticas dos símbolos manipulados, nunca propriedades semânticas. Para um sistema que pretende simular e explicar a inteli-

[9] Esta formulação da tese é devida a HAUGELAND (1989).

gência humana, esta parece uma limitação grande demais. O cognitivismo lida com isto de modo engenhoso, por meio do que CUMMINS (1991) denominou *semântica interpretacional*: computadores são sistemas formais automáticos interpretados.

Imagine-se uma caixa preta para a qual atribui-se a habilidade de jogar xadrez. De que modo é possível confirmar se a caixa possui, de fato, tal capacidade? Em tese, basta observar o comportamento da caixa em um número razoável de situações: pode-se tentar jogar xadrez com ela ou observar partidas com outros jogadores. Se os *outputs* produzidos por ela forem sempre coerentes com as jogadas de xadrez que seriam esperadas nas situações a que ela é submetida, parece não restar motivo para duvidar que ela é capaz de jogar xadrez. Suponha-se agora que os mecanismos internos desta caixa preta sejam explicados sistematicamente, de modo que possam ser descritos nos termos de um sistema formal. Tal sistema explicaria a habilidade que a caixa preta possui apelando apenas às operações formais e aos estados internos formais que ela possui. A habilidade de jogar xadrez, dirá o cognitivista, é a *única* interpretação coerente para este sistema formal.

É possível, claro, supor que ela está na verdade fazendo algo completamente diferente, talvez tentando se comunicar em um idioma alienígena completamente desconhecido. Fosse este o caso, os cientistas estariam interpretando seu comportamento erroneamente, acreditando que seus *outputs* são na verdade jogadas de xadrez. O desafio demandado por esta suposição, contudo, é o de encontrar alguma interpretação alternativa perfeitamente consistente com o comportamento (isto é, com o *output*) exibido. O cognitivista não pode afirmar que encontrar uma alternativa coerente é impossível mas, diante de tal possibilidade, provavelmente se restringiria a desejar boa sorte ao seu desafiante. Ao transpor esta analogia para o caso da mente humana, tem-se um cenário bastante parecido: é preciso encontrar um sistema formal cujo exercício das operações seja coerente com o comportamento inteligente humano. Uma vez encontrado, o resultado será uma descrição dos mecanismos mentais por meio dos quais tal comportamento se dá. Esta tese pode ser sintetizada por meio do que Haugeland batizou de o *mote do formalista*: *"se você tomar conta da sintaxe, a semântica tomará conta de si mesma"* (1989, p. 106).

Diante do exposto, já é sabido que, de acordo com o cognitivismo, é possível automatizar sistemas formais e que estes podem ser interpretados de modo a simular os fenômenos mentais. Contudo, isso parece pressupor que se saiba de antemão quais são os mecanismos que subjazem a mente, ou seja, somente de posse de uma explicação sistemática da mente humana seria possível então simulá-la em uma máquina. Isto é, em parte, verdadeiro. Contudo, a IA não busca apenas simular em máquinas os mecanismos mentais que foram descritos a partir de outras pesquisas (tais como as que são levadas adiante pela psicologia cognitiva). A IA é, ela mesma, uma tentativa de descobrir quais são estes mecanismos. Isso se dá de dois modos. Primeiro, pela formulação de hipóteses acerca de quais são os tipos de símbolos e os conjuntos de operações necessários para simular os mecanismos de algum componente da mente

humana ou algum aspecto do comportamento humano. Segundo, pela formulação de hipóteses acerca do modo como estas operações são efetivamente realizadas. Voltando à analogia com o xadrez: não basta descrever quantas e quais são as peças disponíveis, nem quais são os movimentos possíveis. É preciso também descrever as jogadas que o sistema efetivamente realiza e explicar o motivo de ele ter realizado uma dada jogada e não outra. Este segundo aspecto é realizado por meio de combinações sequenciais de operações denominadas *algoritmos*.

Supondo que o sistema formal disponha de uma operação denominada `soma(x,y)` cujo *output* é a soma de x e y. A realização sucessiva da operação `soma()`, por exemplo, pode ser parte de um algoritmo utilizado para multiplicar um número por outro. Embora não passem de sequencias de operações, algoritmos podem apresentar grande flexibilidade. Uma operação condicional hipotética como `se_x_for_da_cor_y(x,y)` pode ser utilizada para que um certo conjunto de operações seja executado somente no caso de o objeto designado pelo símbolo x possuir uma cor y. Assim, tal como um jogador de xadrez que realiza sua jogada com base no atual estado das peças no tabuleiro, um algoritmo pode resultar em diferentes *outputs*, a depender do estado em que o sistema se encontra. Algoritmos são, neste sentido, uma espécie de receita que especifica a sequência em que determinadas operações devem ser realizadas. São eles que permitem simular os mecanismos da mente humana, permitindo, em tese, que um sistema formal exiba comportamento idêntico quando exposto a situações idênticas às quais a mente humana pode ser exposta. Contudo, se fosse necessário implementar tais algoritmos diretamente em meios físicos, sejam estes eletroeletrônicos, mecânicos ou hidráulicos, dificilmente a IA se mostraria como uma empreitada viável. Seria preciso construir um equipamento físico diferente para cada possível hipótese acerca de quais devem ser os tipos de símbolos, o conjunto de regras e os algoritmos que regem as operações, de modo a simular um aspecto do comportamento humano. Como foi possível evitar esta necessidade?

A resposta vem, novamente, de Alan Turing e de uma versão específica de sua máquina hipotética denominada *máquina universal de Turing* (1936). Tal máquina é um sistema formal que especifica um conjunto de operações e símbolos que permitem *simular* qualquer outra máquina de Turing.[10] O resultado é um sistema formal cujos algoritmos que regem seu comportamento podem ser determinados posteriormente à construção física da máquina, ou seja, trata-se de um computador que pode ser *programado*. Assim, munidos de um único computador programável, os pesquisadores podem elaborar algoritmos e checar se o seu comportamento harmoniza com o comportamento humano que se busca explicar. Viabiliza-se assim um ciclo de desenvolvimento de programas voltados à simulação de diferentes aspectos (isto é, di-

[10] Quando se trata de máquinas de Turing, o termo “operação” é pouco preciso, pois nelas o comportamento do sistema formal é especificado a partir de conjuntos de estados possíveis e regras de transição entre estes estados. Contudo, não é necessário preocupar-se com isto. O ponto importante é deixar claro que conjuntos de operações e conjuntos de regras de transição entre estados são modos diferentes de modelar sistemas formais equivalentes.

ferentes componentes) da mente humana. O modo como estas instruções externas são oferecidas pode variar enormemente, a depender da tecnologia disponível: cartões perfurados, fitas metálicas, *pen drives* etc. O computador "lê" estas instruções e realiza as operações desejadas na sequência especificada. Com efeito, ao invés de se ter uma máquina fisicamente construída para realizar um sistema formal específico, tal como o de um jogo de xadrez, o resultado é a possibilidade de construir uma máquina que pode ser programada para realizar quaisquer operações de qualquer sistema formal.

Diante do cenário apresentado, já é possível sintetizar o que estava por trás do convite de McCarthy em 1956: uma versão embrionária da TCC. Isto se mostra pela presença de três elementos importantes, já apresentados: primeiro, a tese de que explicações sistemáticas fornecem descrições formais adequadas dos componentes funcionais da mente e de suas interrelações. Segundo, a tese de que todo comportamento inteligente pode ser descrito nos termos das operações realizadas por estes componentes funcionais. Em conjunção, estas ideias fornecem a base do cognitivismo. Terceiro, a tese de Turing: para cada sistema formal, existe uma máquina de Turing (i.e. um programa de computador) equivalente. Assim, dada uma descrição suficientemente detalhada de um componente funcional da mente, é possível simulá-la por meio de um computador programável. O mesmo se dá com os processos que resultam em comportamento inteligente. Em adendo, a disponibilidade de computadores programáveis, ainda que de forma escassa, consolidou um cenário de grande (e, sabe-se hoje, exagerado) otimismo quanto ao sucesso da empreitada num curto espaço de tempo. Esta primeira fase da IA costuma ser identificada pela expressão *GOFAI* (*good old fashioned artificial intelligence*),[11] expressão esta que será adotada daqui pra frente quando for necessária uma referência ao conjunto específico das teses aqui expostas.

Até o momento, tomou-se a GOFAI como uma empreitada pautada por um único tema, que é o de descrever os mecanismos por meio dos quais a inteligência humana se mostra. De uma perspectiva panorâmica, essa descrição é pertinente, visto que a GOFAI constitui o cenário amplo em que o FP foi mapeado pela primeira vez. No entanto, os projetos que efetivamente compunham a empreitada expressam uma pluralidade de objetivos específicos. Cabe agora caracterizar melhor o cenário estrito, dentro da GOFAI, no qual o FP se mostrou por primeiro, permitindo compreender melhor seus desdobramentos ao longo dos anos.

Tal cenário específico é constituído por três características importantes: primeiro, havia diferentes graus de preocupação com questões psicológicas na descrição dos processos que constituiriam a inteligência. A GOFAI comportava pelo menos duas abordagens distintas para estudar a inteligência tal como concebida pelo cognitivismo. Esta poderia ser investigada tanto como um fenômeno da mente humana quanto como um fenômeno mais geral. No primeiro caso, as explicações devem apresentar plausi-

[11] O termo foi cunhado por Haugeland (1989).

bilidade psicológica. Uma explicação que precise supor uma capacidade de memória superior à que o ser humano apresenta, por exemplo, não é aceitável. Mesmo em seus primórdios, essa linha de pesquisa se mesclava com o que veio a se tornar a psicologia cognitiva. No segundo caso, não há tais limitações. A inteligência humana é considerada apenas um dos possíveis modos de instanciar uma inteligência. Essa pode, em princípio, manifestar-se de outros modos, inclusive em sistemas cognitivos com muito mais recursos que os disponíveis aos seres humanos. Para esses pesquisadores, essa inteligência ideal geral poderia ser formalizada e expressa na forma de um algoritmo. As limitações da inteligência humanas não eram vistas como características do fenômeno da inteligência ou do algoritmo específico que a descreve, mas sim como limitações do veículo no qual aquele algoritmo fora implementado. Veículos com maior capacidade poderiam apresentar uma inteligência superior, pois permitiriam que o algoritmo da inteligência pudesse ser rodado de modo mais próximo do ideal. O objetivo dessa linha, portanto, era o de descobrir qual seria esse algoritmo da inteligência.

Segundo, havia diferentes formas de descrever estes processos, de acordo com o formalismo adotado. Desde que ROBINSON (1965) desenvolveu um método para construir provas lógicas de modo automático, muitos pesquisadores adotaram a lógica de primeira ordem como o formalismo por meio do qual descreveriam os processos de solução de problemas. De outro modo: a lógica de primeira ordem pode funcionar como a linguagem formal por meio da qual o algoritmo da inteligência pode ser expresso. Isso deu origem à abordagem *dedutivista* da GOFAI. A estratégia utilizada consistia em representar os problemas a ser resolvidos por meio de axiomas. O mecanismo de inferência utilizado para resolvê-los buscava construir provas de teoremas a partir desses axiomas. Encontrar e provar o teorema adequado era encontrar a solução para o problema. Esses axiomas poderiam representar, por exemplo, crenças do sistema cognitivo. Um objetivo ou anseio é um estado de coisas representado por um teorema. Realizar deduções e elaborar uma prova para este teorema é o equivalente a planejar o conjunto de ações necessárias para obter o estado de coisas que o teorema representa. Conforme descreve JANLERT (1987), essa não é uma estratégia que prima pela rapidez. Normalmente, faz-se necessária uma quantidade relativamente grande de processamento (ou seja, é preciso realizar um grande número de operações), mesmo para tarefas simples. Contudo, o ponto forte da abordagem dedutivista estava na sua capacidade de representar e, assim, modelar de forma rica os estados de coisas concebíveis em um dado universo. Em principio, um axioma suficientemente complexo pode descrever qualquer estado de coisas. Considerando que o maior volume de operações pode ser compensado por computadores mais rápidos, a abordagem dedutivista mostrava-se bastante promissora.

Um exemplo de abordagem distinta à dedutivista é a estratégia utilizada pelo GPS (*General Problem Solver*) de Newell e Simon (1959). Trata-se de um programa que objetivava, como o nome sugere, ser capaz de resolver qualquer tipo de problema

que lhe pudesse ser dado como *input*. No entanto, ao invés de guiar-se pelo uso da lógica para estudar a inteligência como fenômeno mais amplo, fazia-se uso de uma estratégia inspirada no modo como o ser humano raciocina, a fim de estudar a inteligência humana em particular. O processo de solução de problemas utilizado era predominantemente heurístico. Tais heurísticas eram compreendidas como uma espécie de "regra de ouro" a ser aplicada em cenários específicos. Saltos lógicos baseados em elementos que um cenário apresenta tipicamente são exemplos comuns. Além de heurísticas, o programa incluía também processos que tentavam "adivinhar" o resultado com base em informações incompletas, e mesmo processos de tentativa e erro. Intuitivamente, abordagens heurísticas parecem muito mais próximas do modo como o ser humano raciocina quando tentando resolver algum problema. Contudo, para o dedutivista, este apelo à intuição pode ser enganoso. Ele crê que a lógica é uma ferramenta muito mais poderosa para a simulação dos processos da inteligência. Contemporaneamente, esse parece um debate um tanto datado. Contudo, como se poderá perceber ao longo da discussão, boa parte das questões envolvendo o uso de formalismos e estratégias de solução de problemas continuam presentes na IA contemporânea.

O terceiro e último ponto importante na caracterização do cenário específico em que o FP surgiu diz respeito ao modo como as pesquisas eram realizadas. O uso de *micro-universos* era (e ainda é) praticamente uma constante. Um micro-universo pode ser um recorte do mundo real, tal como a situação específica de estar em um restaurante. Assim, para o sistema cognitivo, nada além do que era explicitamente descrito como parte desse recorte precisava ser considerado. Outra versão comum dos micro-universos são cenários artificiais criados sob medida para certos programas. Talvez o exemplo mais conhecido seja o do "mundo dos blocos" do *SHRDLU*, criado por WINOGRAD (1972). Trata-se de um cenário onde tudo o que existe são blocos de diversas formas sobre uma superfície e um braço mecânico capaz de manipulá-los. Winograd fez uso deste micro-universo para demonstrar a possibilidade de um computador "compreender" a linguagem natural (as ordens ao braço mecânico eram passadas em inglês).

Há pelo menos duas razões importantes por trás da opção pelo uso de micro-universos: primeiro, as limitações práticas das ferramentas de trabalho, isto é, dos computadores. A capacidade de processamento e o volume de informações com que se conseguia trabalhar era relativamente reduzido. Segundo, a simplificação da pesquisa: cenários mais simples permitem modelos mais simples. O uso do mundo dos blocos permitiu a Winograd realizar um programa capaz de compreender comandos em inglês e responder perguntas sobre seu próprio comportamento também em inglês. Diante de uma pergunta como "por que você moveu o bloco azul?", o SHRDLU poderia responder "porque eu precisava liberar o acesso ao bloco amarelo sob ele". Contudo, fora do micro-universo para o qual havia sido elaborado, o programa não era capaz de compreender sentenças ou responder de forma apropriada. Micro-universos

são utilizados na esperança de que princípios gerais possam ser capturados a partir de soluções algorítmicas elaboradas para estes cenários específicos. Uma vez descobertos, estes poderiam ser generalizados para outros cenários. A complexidade dos casos específicos aumenta, mas o princípio permanece. O uso desta estratégia na IA é, por vezes, considerada um caso da chamada *falácia do primeiro passo*. Dreyfus descreve esta falácia da seguinte forma: "*é como afirmar que o primeiro macaco que subiu em uma árvore estava fazendo progresso no objetivo de voar até a lua*." (2007, p. 68). Se a IA conseguiu descobrir algum princípio da inteligência geral, esta é uma questão que permanece em aberto. O que se sabe por certo é que ela conseguiu encontrar problemas e dificuldades que parecem muito mais gerais, a ponto de emergir em todas as abordagens, dedutivistas ou não, focadas na inteligência geral ou não. Como se verá, este é o caso do FP.

Uma vez que se tenha apresentado as características relevantes, é possível sintetizar o cenário específico do seguinte modo: o FP surgiu na tentativa de modelar cenários (micro-universos) por meio de lógica de primeira ordem no âmbito da abordagem dedutivista, enquanto se buscava estudar a inteligência como um fenômeno geral, tendo as teses que caracterizam a GOFAI como pano de fundo. Manter este cenário em mente ajudará a compreender a grande quantidade de diferentes descrições do FP presentes na literatura que orbita a GOFAI e os debates filosóficos que daí advieram. Boa parte dos diversos modos de compreensão do FP, bem como da confusão em torno de sua real natureza, pode ser atribuída ao escopo em que cada autor insere sua análise. Isto deu origem a pontos de vista completamente opostos. Como se verá adiante, autores como McDermott (1987) acreditam que o FP é um problema exclusivo da abordagem dedutivista. No outro extremo, Dreyfus (2007) especula que o problema é fruto de uma limitação embutida no uso de modelos computacionais, e que estes devem ser completamente abandonados. Enquanto um autor quer apenas corrigir a rota, o outro sugere que o navio inteiro deve ser sacrificado. Compreensões tão distintas não se dão, contudo, apenas em virtude da pluralidade de escopos nos quais o FP foi analisado. É igualmente importante saber distingui-lo de outros problemas relacionados, mas distintos. Introduzir o ferramental necessário para realizar esta distinção é o objetivo da próxima sessão.

1.2 Problemas de relevância

O FP se insere numa classe de problemas que pode ser denominada *problemas de relevância*. O que caracteriza este tipo de problema no cenário da GOFAI? Ele surge em diversas ocasiões, mas as versões mais interessantes e desafiadoras se mostram quando um algoritmo precisa simular a flexibilidade do comportamento inteligente humano em função dos diferentes contextos em que ela pode se apresentar.[12]

[12] Note-se a importância de manter em mente as distinções previamente apresentadas entre pesquisas no interior da GOFAI. Uma solução pode ou não ser aceitável, a depender do objetivo específico da

O adversário fez sua jogada no xadrez, como responder? Alguém tocou a campainha, o que fazer? Intuitivamente, estas perguntas parecem demandar respostas que se originam de habilidades bastante distintas. É relativamente fácil compreender de que modo uma resposta adequada no xadrez pode ser elaborada por meio de operações e cálculos que partem do atual estado das peças no tabuleiro. Contudo, no caso da campainha, torna-se mais difícil enxergar que tipo de operação ou cálculo poderia fornecer a resposta adequada ao evento.

Não obstante, a GOFAI busca explicar os dois comportamentos por meio de uma mesma estratégia: a mente recebe um *input* (o tocar da campainha ou a jogada do adversário), processa este input por meio de seus mecanismos internos e gera um *output*, uma resposta. No caso do xadrez, a resposta pode ser tanto uma jogada possível quanto a conclusão de que este é o momento de pedir ao adversário que se decrete um empate. Por sua vez, o tocar da campainha pode gerar a crença de que há alguém do outro lado da porta. Qual o problema, então? O xadrez, sendo um jogo, tem um *objetivo* fixo e bem definido que vale para qualquer partida: o objetivo último de toda jogada é dar um passo rumo a vitória. Este objetivo funciona como critério para que se possa selecionar, dentre todas as jogadas possíveis (isto é, as jogadas válidas dado o atual estado do tabuleiro), aquela que melhor se aplica ao cenário presente. Contudo, o caso da campainha não parece dispor de um critério pré-definido de modo tão rígido. O comportamento inteligente humano não se parece pautar por um critério único e fixo de modo que, sempre que uma campainha tocar, a mente humana tenha condições de guiar suas operações por meio dele. Assim, processar a reação adequada ao estímulo (tocar a campainha) parece depender de uma noção por ora obscura do que é ou não “adequado” a depender do contexto em que o estímulo se deu.

O cognitivista poderia responder de imediato: assim como o contexto de uma partida de xadrez equivale ao estado atual das peças no tabuleiro, um contexto cognitivo nada mais é do que um certo estado informacional do sistema que a mente humana é. Se o estado for escrito em suficiência de detalhes, não há porque duvidar que seja possível descrever os mecanismos por meio dos quais a reação adequada possa ser processada. No entanto, há uma dificuldade com esta resposta que salta aos olhos: quantos e quais são os possíveis estados do sistema, isto é, quantos e quais são os possíveis contextos em que o sistema pode se encontrar? Mesmo no caso do xadrez, o número é gigantesco, mas finito. Por outro lado, o número de contextos em que a mente humana pode operar parece ter caráter indeterminado e infinitário, no mesmo sentido em que o número de frases possíveis para um idioma é potencialmente infinita. Qual a reação adequada à campainha? Pode-se abrir a porta, caso seja um contexto em que se está a esperar alguém. Pode-se também concluir que o melhor é perguntar de quem se trata antes de abrir a porta. Da mesma forma, gritar pode ser uma reação

---

pesquisa no interior da qual esse desafio se ponha.

adequada caso se esteja escondido em uma casa sob um serviço de proteção a testemunhas. É plausível supor que todas essas possíveis situações sejam explicitamente consideradas como estados possíveis de um sistema formal? Não é o momento de se aprofundar nestas questões. Este é um esboço superficial de um debate que será detalhado nas seções subsequentes. Por ora, o importante é notar que a dificuldade não está em processar *inputs* e gerar *outputs* possíveis. O desafio peculiar posto pelos problemas de relevância está em identificar, dentre todos os *outputs* possíveis, aqueles que devem ser levados em conta no contexto em que o sistema cognitivo se encontra, isto é, aqueles que são relevantes.

O motivo pelo qual essa classe de problemas é especialmente desafiadora para a GOFAI não é imediatamente intuitivo. Esses podem ser facilmente confundidos com problemas menores, ou mesmo com meros detalhes de implementação que podem ser resolvidos num âmbito puramente técnico. Não raro, são também tomados como fruto de limitações tecnológicas que podem vir a ser resolvidos ou ignorados a partir de computadores mais rápidos ou com maior quantidade de memória. Essas tentativas de caracterizar problemas de relevância nesses termos não é infundada. Alguns problemas menores, perdidos nas minúcias das pesquisas realizadas, foram reconhecidos e resolvidos deste modo. Contudo, alguns deles, em especial o FP, têm o odioso hábito de voltar pela janela uma vez que tenham sido retirados da sala pela porta. Em geral, isso acontece quando uma tentativa de solucioná-lo revela-se uma mera reformulação do problema. Em vez de resolvê-lo, apenas desloca-se seu ponto essencial. A dificuldade é análoga à que se dá em certas narrativas mitológicas que tentavam explicar o fato de a terra não cair no espaço por estar apoiada sobre quatro elefantes. Evidentemente, o problema retornava: o que sustenta os elefantes? Na versão mais conhecida do mito, estes são sustentados por uma tartaruga gigante. Contudo, a pergunta retorna: o que sustenta a tartaruga? A história do FP tem estrutura semelhante. Por vezes ele foi dado como resolvido, mas apenas para logo ressurgir sob nova formulação. Por isso, compreender o real desafio que o FP constitui demanda conhecer, ainda que panoramicamente, sua história, algumas das tentativas de solução apresentadas e as razões pelas quais estas falharam.

1.2.1 O *frame problem* na Inteligência Artificial

Conforme visto, o FP nasceu em uma linha de pesquisa da GOFAI pautada pelo uso de micro-universos, pela lógica de primeira ordem e pelo objetivo de estudar a inteligência de modo não psicologicamente restrito. A fim de caracterizar melhor o modo como foi por primeiro descrito, ele será apresentado tal como se manifestaria em um micro-universo inspirado no “mundo dos blocos” da obra de Winograd: DR31FUS é um robô construído sob medida para habitar um universo também construído sob medida. Seu corpo físico é composto por nada além de um braço mecânico com uma garra que lhe permite trocar a posição de objetos simples. Seu universo, isto é, tudo o

que existe para ele, é uma pequena sala fechada com quatro blocos retangulares coloridos: o bloco azul TOD, o bloco amarelo HEI, o bloco verde MCD e o bloco vermelho MPT.

Figura 1 – Duas situações do micro-universo de DR31FUS

DR31FUS consegue movimentar os blocos e colocá-los em qualquer ponto da sala, inclusive uns sobre os outros. Tome-se agora uma situação S0 tal em que os blocos se encontrem dispostos do seguinte modo: TOD e MCD estão no chão da sala em cantos opostos, HEI está sobre TOD e MPT está sobre HEI (vide S0 na fig. 1). DR31FUS mantém uma representação desta situação em sua memória e por isso pode responder perguntas sobre ela, informando, por exemplo a localização ou a cor de algum bloco. Em dado momento, ele “decide” alterar a situação atual, colocando todos os blocos sobre MCD. Para isso, ele precisa realizar uma série de operações: calcular se há espaço disponível no chão para deixar algum bloco ali temporariamente, colocar MPT no chão para liberar HEI, colocar HEI em algum lugar no chão para liberar TOD e só então colocar TOD sobre MCD. Após isto, ele poderá colocar novamente HEI sobre TOD e, finalmente, MPT sobre HEI, gerando assim uma única pilha com os quatro blocos, uma situação denominada S6 (vide S6 na fig. 1).

Pode-se supor que agora DR31FUS tenha uma representação de S6 em sua memória e pode, tal como antes, responder perguntas sobre esta nova situação. Essa expectativa é quebrada, porém, se logo após o último movimento realizado (colocar MPT sobre HEI), ele for convocado a responder à seguinte pergunta: qual a cor de MPT? De modo um tanto surpreendente, DR31FUS não sabe mais responder esta pergunta. Trata-se de uma pergunta tão simples, cuja resposta parece tão óbvia, que é difícil de acreditar que, diante dela, DR31FUS não saiba mais o que dizer. Alterar um bloco de lugar não altera sua cor, que permaneceu rigorosamente a mesma durante a passagem de uma situação para outra. Como é possível que a informação tenha sido “perdida”? Na verdade, a informação não foi perdida. O que se perdeu foi a capacidade utilizá-la. Isto é possível se DR31FUS sofrer do FP.

Esse problema foi primeiro descrito e nomeado por MCCARTHY; HAYES (1969). Contudo, nenhum deles tinha a pretensão ou interesse em descrever algo tão amplo quanto o que ele veio se tornar. O que tinham em mente era um problema de caráter

lógico muito específico. McCarthy e Hayes eram partidários da abordagem dedutivista que utilizava lógica clássica de primeira ordem. Em particular, Hayes era um ferrenho defensor da lógica como o formalismo mais adequado para tratar da inteligência.[13] Contudo, uma característica essencial da lógica clássica coloca-se de imediato como desafio: a atemporalidade. Se uma dada sentença é verdadeira em um dado momento, ela sempre o será. Esta não é uma característica desejável de nenhum sistema formal por meio do qual se pretenda representar um universo em que ocorrem mudanças.

Para lidar com esta dificuldade, Hayes e McCarthy desenvolveram uma extensão da lógica clássica denominada *cálculo de situação* (*situation calculus*). Trata-se de uma lógica polissortida (*many-sorted*) que especifica pelo menos três diferentes tipos lógicos: situações, fluentes e ações.[14] Uma situação é *"... o estado completo do universo em um dado instante de tempo."*, ou seja, é uma espécie de descrição de todos os fatos de um modelo parcial do universo num dado momento (MCCARTHY; HAYES, 1969, p. 18). Em micro-universos artificiais ou em recortes arbitrários do universo real, esta noção parece não apresentar problemas. No caso do mundo de DR31FUS, o universo em questão é constituído basicamente por blocos de diferentes cores distribuídos sobre uma superfície determinada. Isso permite que uma situação seja representada integralmente por poucos axiomas. Porém, se o universo em questão for tão grande quanto se supõe ser o universo real, então é evidentemente impossível representá-lo na sua totalidade. Nesse caso, uma situação compreenderia apenas os aspectos do universo de algum modo acessíveis ao agente.

Os fatos que constituem as situações são representados com a ajuda de *fluentes*. Um fluente é uma espécie de variável cujo valor pode ser alterado no passar de uma situação para outra. A localização de objetos, bem como suas propriedades, são exemplos de elementos representados por fluentes. Os valores dos fluentes só podem ser alterados por *ações*. Quando realizadas, ações levam o universo a uma nova situação. Assim, os blocos de DR31FUS são representados por fluentes, e a ação `mover_objeto` é responsável por levar o universo de uma situação a outra. Essencialmente, a ideia é descrever a situação atual do universo por meio de um conjunto de axiomas e fluentes. Quando uma ação é realizada, métodos de cálculo de prova são empregados para calcular todos os teoremas que descrevem a situação resultante. Assim, somente a realização de uma ação faz com que se dê o salto de uma situação para outra, e o único elemento que pode realizar uma ação é o próprio sistema. Além disso, não é possível realizar ações concomitantes, mas somente ações encadeadas, uma após a outra. Consequentemente, o sistema era determinístico: tudo o que constitui uma nova situação (isto é, todos os efeitos de uma dada ação) deve ser calculável a partir do conjunto de axiomas e fluentes da situação anterior.

Como visto, uma *ação* é um tipo de operação que pode alterar os valores de flu-

[13] Ver HAYES (1977) para uma defesa dessa tese.

[14] Conforme SHANAHAN (1997). A descrição aqui apresentada é extremamente simplificada a fim de evitar detalhes técnicos não relevantes para a presente discussão.

entes e estabelece um novo conjunto de axiomas que descrevem o universo, ou seja, uma nova situação. As ações descrevem, portanto, as leis de movimento, a "física" do universo. Toda e qualquer mudança neste universo é, necessariamente, fruto de uma ação do agente. Para que uma ação possa desempenhar este papel, ela precisa ser definida por meio de um conjunto de axiomas. São os chamados *axiomas de efeito* (*effect axioms*). Tais axiomas explicitam os resultados obtidos pela ação, de modo que o sistema possa sempre recalcular os valores que os fluentes terão na situação resultante. Uma ação denominada `mover_objeto(x,y)`, por exemplo, tem como efeito que, na situação resultante, é verdadeiro que o objeto *x* está na posição *y*. Intuitivamente, descrever os efeitos de uma ação parece ser condição suficiente para que o sistema possa recalcular os valores de todos os fluentes de maneira adequada. Ao realizar uma ação e saltar de uma situação à outra, o sistema precisa apenas calcular o valor dos fluentes afetados pela ação (a nova localização de um bloco, por exemplo) a partir dos axiomas de efeito, e todo o resto permanecerá inalterado.

Infelizmente, a lógica clássica não respeita essa intuição. Como a situação é um tipo lógico distinto, não sendo ela mesma uma sentença no conjunto de axiomas a partir dos quais o sistema realiza os cálculos, é preciso especificar explicitamente, para cada ação, não apenas os efeitos produzidos, mas também aquilo que não é afetado de modo algum. As consequências deriváveis dos axiomas que constituem uma situação S0 não estão integralmente disponíveis em S6, a menos que estejam explicitadas como não efeitos da ação que levou a S6. Aos axiomas que explicitam os não efeitos de uma dada ação, é dado o nome de *axioma de frame*. Assim, uma ação como `mover_objeto(x,y)` é descrita não apenas pelos seus axiomas de efeito, mas também pelos seus axiomas de *frame*. Além de explicitar que na situação resultante o objeto *x* estará na posição *y*, é preciso também especificar, por exemplo, que sua cor, seu tamanho ou sua forma permanecem inalterados. Sem axiomas de *frame*, os fluentes que não são afetados pela ação não podem ser recalculados, figurando na situação resultante somente como indeterminados ou como possibilidades (ou seja, não podem ser provados). O resultado é um comportamento tal qual o apresentado por DR31FUS no exemplo acima: após a execução de uma ação, seu modelo interno do universo não lhe permite mais realizar as inferências necessárias sequer para afirmar que um objeto vermelho permanece vermelho.

No entanto, a solução para o caso de DR31FUS parece relativamente simples: seu micro-universo contém apenas uma ação (mover um objeto) e um tipo de objeto (bloco) com duas propriedades: cor e localização. Basta, portanto, que sejam acrescentados dois axiomas de *frame* à definição da ação: o primeiro especifica que a cor do objeto permanece inalterada; o segundo, por sua vez, especifica que a localização de todos os demais objetos, exceto daquele que é movido, também permanece inalterada. O problema com esta abordagem salta aos olhos de maneira quase imediata: descrever tais axiomas para um micro-universo tão simples parece trivial, mas o que acontece caso se esteja a descrever ações para um universo mais complexo, com

diversos tipos de objetos e propriedades? De fato, em um domínio com M ações e N propriedades, o número de axiomas de *frame* necessários será de MN. Em um universo suficientemente complexo (como presumivelmente é o do caso universo real), o número de axiomas necessário oscilará entre o gigantismo e o infinito. Surge assim o *Frame Problem*, cujo nome deriva do problema de especificar axiomas de *frame* para determinar os não efeitos de ações em universos mais complexos que o de DR31FUS.

Resta agora ver a razão pela qual o FP é considerado um problema de relevância. No contexto descrito, o FP apresenta dois aspectos. Primeiro, o *aspecto inferencial.*[15] Diante de uma ação como mudar o bloco de lugar, o sistema deve ser capaz de inferir eficientemente se aquilo que era verdadeiro antes permanece verdadeiro. A solução cognitivista para este aspecto do problema é algorítmica, embora possa incluir estratégias como o uso de heurística e outros recursos. Contudo, calcular uma prova que considera MN axiomas de *frame* pode mostrar-se problemático ou impeditivo, tanto em termos de consumo de recursos computacionais, tais como memória, quanto em tempo de processamento. Um computador precisa ser capaz de lidar não apenas com um gigantesco volume de axiomas de *frame*, mas também com as etapas provisórias dos processos de cálculo. Assim como um ser humano precisa "armazenar" temporariamente, seja na memória, seja no papel, os passos intermediários utilizados na solução de um problema matemático, o computador realiza algo parecido para cada problema. Se o computador não souber distinguir as inferências relevantes das irrelevantes, isto significa que ele precisará também armazenar temporariamente os passos intermediários de cada possível inferência a partir de cada axioma e de cada possível combinação deles, bem como de cada possível inferência a partir de cada uma destas inferências, e assim por diante, até esgotar as inferências possibilitadas por aquele estado de coisas no universo. A disponibilidade de recursos computacionais para isto é menos preocupante porque sua disponibilidade pode variar, conforme a tecnologia disponível.

Dois exemplos de recursos são de particular importância: a capacidade de memória e as tecnologias que permitem armazenar dados de forma compacta. O número de *bits* (isto é, símbolos) que podem ser armazenados e disponibilizados para processamento é hoje algumas milhares de vezes superior ao que se conseguia armazenar nos primeiros computadores. Em paralelo, diferentes modelos de armazenamento permitem reduzir o espaço necessário para armazenar um dado conteúdo. O advento do formato MP3 é um exemplo disso: uma mesma música passou a ser armazenável em 1/10 ou 1/15 do espaço que era antes necessário, e isto sem perda de conteúdo significativo. Não bastasse, é possível também reunir conjuntos de computadores, permitindo que as capacidades de armazenamento de cada um sejam somadas, gerando

[15] Autores como MCDERMOTT (1987) utilizam o termo "aspecto computacional" para designar esta característica. Contudo, como se verá, a concepção de FP aqui exposta diverge da concepção de McDermott. Além disso, McDermott tem em mente um escopo muito específico, que é o da GOFAI, e a versão aqui apresentada pretende ser generalizável a escopos mais amplos, daí a opção por um termo mais vago e abrangente.

os chamados supercomputadores. Assim, mesmo diante da grande demanda, é pouco provável que o impedimento à uma solução pautada por um grande número de axiomas de *frame* venha da indisponibilidade de tecnologia para construir um computador com capacidade de armazenamento suficientemente grande.

Por outro lado, o tempo disponível para realizar as operações sempre será uma questão. Não importa quão rápido um computador execute uma determinada operação (mesmo considerando a velocidade estrondosa dos ainda hipotéticos computadores quânticos), para universos suficientemente complexos, a quantidade de cálculos necessários para calcular provas envolvendo MN axiomas pode revelar-se grande demais para ser realizável num período de tempo razoável.

Não haveria tanta dificuldade em aceitar esta consequência se o problema alvo fosse realmente difícil, inclusive para seres humanos. No entanto, não se está a tentar explicar o modo como a inteligência lida com problemas extremamente complicados, mas sim o modo como ela lida com demandas simples que ocorrem em virtualmente todo e qualquer exercício de suas capacidades. A reação ao tocar da campainha depende da capacidade de distinguir de modo eficiente o que é ou não afetado por este evento. Além disso, parece difícil aceitar que, ao tentar manter atualizada a representação que possui do cenário ao seu redor, a mente deva realizar de modo incrivelmente rápido um grande número de inferências irrelevantes apenas para, em seguida, descartá-las: mover o bloco não altera sua cor; mover o bloco não altera sua forma; mover o bloco não altera a crença de que choveu no dia anterior, e assim por diante. Mesmo hoje, muitos desses cálculos demandariam horas, dias ou anos de processamento por parte dos computadores disponíveis.

Não obstante, é pouco provável que a dificuldade esteja apenas na capacidade de realizar grande número de cálculos numa velocidade gigantesca. É mais razoável supor que haja algum problema com o algoritmo utilizado: ele não simula os processos da inteligência de modo adequado porque não explica como é possível distinguir inferências relevantes das irrelevantes sem ter que realizá-las. Afinal, para que o problema se mostre, basta que exista a necessidade de realizar uma inferência para só então considerá-la irrelevante, pois segue-se daí que o sistema precisará analisar todo o universo para determinar se uma inferência realizada é ou não relevante. Isso faz do FP um problema de relevância. O aspecto inferencial do FP é, basicamente, a dificuldade de explicar como o sistema realiza somente as inferências relevantes de um modo eficiente.

Segundo, e talvez mais importante, há o caráter *ad hoc* dos axiomas de *frame*. Trata-se do modo como o problema se manifesta para o *designer* do sistema cognitivo. Uma solução adequada para este aspecto do FP não envolve um algoritmo, mas um modelo. Como seria possível especificar por princípio os efeitos e os não efeitos de uma ação? Mesmo ações simples como movimentar objetos podem resultar em efeitos inesperados ou indesejados. Movimentar um pires, por exemplo, pode fazer com que a xícara o acompanhe. Se isto não for levado em conta, a representação interna

do sistema pode conter absurdos como o de uma xícara que “flutua” sobre a mesa. Por outro lado, nem sempre um pires leva consigo uma xícara. Lidar adequadamente com um universo complexo parece demandar axiomas de *frame* que podem valer em alguns contextos, mas não em todos. Isso aumenta o desafio, uma vez que seria preciso mapear previamente todos os possíveis contextos de realização das ações do sistema. Sem isso, o axioma parece colocado ali de modo *ad hoc* para resolver uma dificuldade específica, e não um problema geral. Não por acaso, este aspecto do FP é o principal responsável pelo interesse filosófico que despertou. Já é possível notar, aliás, como a dificuldade supera o escopo em que o problema inicialmente surgiu: ele se mostra um desafio tanto para a emulação de uma inteligência idealizada, quanto para quem quer descrever o tipo de atividade que se passa na maquinaria interna da mente humana. Se para quaisquer destes projetos for necessário formular um conjunto infindável de axiomas apenas para explicar o modo como se lida com efeitos e não efeitos de ações, então a GOFAI não parecer ter um caminho promissor diante de si.

SHANAHAN (2016) condensou estes aspectos no que denominou *lei da inércia do senso comum* (LISC). Cotidianamente, todo ser humano parece capaz de usar o senso comum para determinar os efeitos e os não efeitos de suas ações. Ainda que de modo falível, é possível saber quando é ou não adequado esperar por certos efeitos. Via de regra, as pessoas sabem que mudar a localização de um objeto não altera sua cor e que esta possibilidade pode ser tomada como irrelevante na maioria absoluta dos casos. Nesses termos, o problema parece ser que todo ser humano é capaz de fazer uso de algo equivalente à LISC para determinar o que é ou não relevante, mas ninguém parece saber como é possível *modelá-la*. Ao invés de uma infinidade de regras, o senso comum pode incluir uma cláusula *ceteris paribus* cuja aplicação varia conforme o contexto, por exemplo. Mas uma tal cláusula não pode ser representada como apenas mais um axioma no sistema, pois isto a tornaria parte do problema: em quais cenários ela seria relevante? Como lidar com isso? Seria a LISC composta por uma enorme quantidade de axiomas que mapeiam efeitos e não-efeitos? Seria ela um modo *default* de raciocínio? Caso fosse possível ensiná-la a alguém, qual seria o conteúdo a transmitir? O FP tal como manifesto enquanto problema técnico pode ser sintetizado como o desafio de modelar esta cláusula e habilitar o sistema a fazer uso dela.

1.3 Primeiras tentativas de solução

Dado o contexto restrito em que o FP primeiro surgiu, é legítimo questionar se o problema não poderia originar-se de alguma característica da abordagem dedutivista ou, de modo ainda mais específico, do uso de cálculo de situação ou da lógica de primeira ordem. Como será visto em mais detalhe adiante, houve, entre os pesqui-

sadores da GOFAI, quem acreditasse nisso e sugerisse que o uso de formalismos e estratégias distintas seriam capazes de evitar o surgimento do FP. Contudo, as diversas tentativas de solucioná-lo, bem como seus recorrentes fracassos, sugerem que o FP tem natureza distinta.

Segundo HAUGELAND (1987), as diversas tentativas de solução podem ser divididas em dois tipos de estratégias: *cheap test* e *sleeping dog*. Ambas tentam, por diferentes meios, modelar a LISC pela circunscrição de subconjuntos de possíveis não efeitos de ações, ou seja, tentam garantir que somente os cálculos mais *relevantes* sejam realizados. A principal diferença entre as duas abordagens está no modo como a circunscrição é realizada.

1.3.1 Estratégias *cheap test*

Haugeland descreve estratégias do tipo *cheap test* como aquelas que *"... examinam toda a situação (tal como representada) rapidamente, para dizer de relance a maior parte do que é irrelevante para um certo evento e não será afetado"* (1989, p. 205). Elas se inspiram de modo direto em uma observação empírica simples: as pessoas parecem "saber" os conjuntos de efeitos que são comuns e os que são incomuns, raros ou excepcionais, ou seja, as pessoas sabem que movimentar um objeto, em geral, não altera sua cor ou sua forma. Esse conhecimento é então modelado na forma de categorizações prévias de eventos do mundo. A estratégia de categorização pode variar. Duas delas serão rapidamente apresentadas aqui: a primeira busca categorizar os efeitos por meio da interação entre as propriedades dos objetos. A segunda o faz pela interação entre os próprios objetos. Esses dois critérios de forma alguma esgotam as propostas *cheap test* existentes na literatura que orbita a GOFAI, mas fornecem material suficiente para compreender os atrativos e os limites da estratégia.

Uma versão da proposta de categorizar os eventos do mundo a partir das relações entre as propriedades de objetos foi esboçada por Hayes no mesmo artigo em que o FP foi apresentado (1969) e também em artigo posterior dedicado ao tema (1973). Ao descrever as propriedades dos objetos do universo, estas são tipificadas. Trata-se, essencialmente, de realizar uma categorização prévia dos tipos de propriedades presentes no universo em questão, tais como massa, cor, posição em um plano cartesiano, etc. Propriedades ligadas à localização do objeto podem constituir um tipo. Propriedades ligadas à sua forma ou cor podem constituir outro, e assim por diante. Uma vez categorizadas, faz-se um mapeamento prévio das relações entre esses tipos de propriedades. Isso permite categorizar o modo como diferentes eventos as afetam. As propriedades ligadas à localização do objeto, por exemplo, não afetam as propriedades ligadas à sua forma ou cor. Assim, há uma drástica redução no número de relações a modelar e de cálculos a realizar. Essa é uma forma de dividir o mundo em *blocos*.[16] Ações podem ser igualmente categorizadas a partir dos blocos que afe-

[16] Hayes, confusamente, chamou estes blocos de *frames*. Contudo, um *frame* compreendido como um

tam, isto é, dos tipos de propriedades que afetam. Uma vez mapeados os blocos que podem ser afetados pela ação, somente os fatos contidos naqueles blocos são considerados. Qualquer possível efeito que não conste explicitamente na definição dos blocos que uma determinada ação afeta pode ser ignorado. O universo não se altera exceto por meio daquilo que fora previamente mapeado.

À primeira vista, isto parece lidar tanto com o aspecto *ad hoc* quanto com o aspecto *inferencial* do FP: ao invés de formular incontáveis axiomas de *frame*, elabora-se um conjunto reduzido de blocos. Viabiliza-se assim o trabalho do *designer* do sistema cognitivo e do próprio sistema cognitivo, pois este ganha um critério que lhe permite circunscrever as derivações que deve ou não considerar, sem testá-las uma a uma. Voltando ao exemplo de DR31FUS, pode-se supor um dado axioma A que descreve a caixa MCD como verde em uma situação S0 e uma ação que altere a localização desta caixa, levando a uma situação S1. O axioma pertence a um dado bloco do mundo (o bloco que se preocupa com as cores) e a ação afeta somente outro bloco (o bloco com as propriedades relativas à localização). É possível então inferir com segurança que A permanece verdadeiro após a ação checando apenas os blocos a que cada propriedade pertence.

Embora não façam uso de axiomas de *frame*, abordagens *cheap test* também buscam modelar a LISC como um conjunto de regras que guiam as operações do sistema cognitivo. Elas podem constituir soluções para o FP se conseguirem elaborar critérios flexíveis (a fim de lidar com o aspecto *ad hoc* do FP) e que reduzam a quantidade de testes necessários a um número computacionalmente viável (lidando assim com o aspecto *inferencial* do FP). Porém, as dificuldades desta abordagem se mostram rapidamente: quão refinados devem ser os blocos que dividem o mundo? Qual deve ser a estratégia? Poucos blocos com grandes quantidades de fatos ou muitos blocos com menos fatos a tratar? Essas perguntas estão relacionadas à uma questão fundamental: como lidar com ações que afetam elementos em mais de um bloco? A importância dessa decisão é devida à importância de se saber quantas e quais serão as ações disponibilizadas no sistema. A depender do modo como essas perguntas são respondidas, o problema original pode retornar pela janela, agora reformulado como a dificuldade em lidar com grande quantidade de blocos e suas relações de interdependência.

Assim, quanto mais precisos ou ricos forem os critérios de demarcação dos blocos, maior será a quantidade de testes necessários para determinar efeitos e não efeitos, e mais complexos estes testes serão. Isto coloca em xeque o cerne da estratégia, pois há o risco de não se ter qualquer ganho real ao trocar a necessidade de especificar sucintamente axiomas de *frame* pela necessidade de uma especificação

---

bloco e um axioma de *frame* são coisas distintas. O primeiro aponta para um critério de tipificação de eventos. O segundo, para o papel que um determinado axioma cumpre num sistema cognitivo, conforme já explicitado. A fim de evitar confusão, optou-se por ignorar este uso que Hayes faz do termo “frame” no sentido de bloco e usar apenas o termo “bloco” para referir-se ao conceito apresentado.

sucinta de blocos. Por outro lado, mapeamentos pouco precisos podem levar a um número excessivo de falhas, que é o problema inicial para o qual os blocos seriam uma solução. Embora ainda seja possível argumentar que esta versão do problema tenha sido reduzida à uma dificuldade de implementação (basta achar um equilíbrio adequado entre ações e blocos), a tentativa de implementá-la faz emergir dificuldades maiores.

Tome-se a ação `mover_objeto()` como exemplo. Ela afeta apenas o bloco "localização" e não o bloco "cor". Contudo, não é possível mapear previamente todas as propriedades do objeto nas quais a ação incorrerá sobre. Se o objeto movido for um camaleão, ele poderá ou não ter sua cor alterada como efeito da movimentação (a depender das circunstâncias do ambiente, pois ele pode ainda ser movido para um local que não o faça alterar sua cor). Assim, a tipificação resultante da divisão do mundo em blocos parece, ela mesma, mostrar-se sensível às circunstâncias. Na ausência dessa sensibilidade, permanece problema de inferir, caso a caso, se uma ação afetou ou não a cor do objeto. Tal como no caso dos axiomas de *frame*, a classificação em blocos só é efetiva em cenários previamente circunscritos, tais como o micro-universo de DR31FUS.

Fora destes micro-universos, seria preciso classificar previamente também os diferentes cenários em que o sistema cognitivo se encontra. De posse desta classificação, o sistema cognitivo poderia então selecionar os blocos adequados ao cenário específico e, num terceiro passo, selecionar o conjunto de testes que deve realizar. Tanto MINSKY (1997) quanto SCHANK (1975) buscaram viabilizar a tipificação destes cenários por meio de *estereótipos*.[17] Assim, é possível mapear cenários como o de "estar em uma floresta" e o de "estar em uma sala branca vazia". No primeiro cenário, a ação `mover_objeto` aplicada ao camaleão afetaria os fatos contidos no bloco que se preocupa com a cor do objeto, já no segundo cenário, não. Evidentemente, tudo isso precisaria ser previamente categorizado numa descrição estereotípica de cada uma das situações. É preciso que o sistema cognitivo conheça estereótipos de situações como "estar em um restaurante", "estar no cinema" ou "estar em sala de aula".

Porém, o uso de estereótipos traz consigo alguns desafios adicionais: primeiro, uma vez que o sistema compreenda a si mesmo como estando em uma situação representada pelo estereótipo X, como reorganizar o conjunto de axiomas que representam

[17] Para falar em estereótipos, Schank faz uso do termo "script". Minsky, por sua vez, faz uso do termo "frame", contribuindo ainda mais para a confusão frequente na literatura em torno do termo. Não bastasse o uso do termo para designar um tipo de axioma (axiomas de *frame*) e o uso para designar uma categorização de propriedades do mundo (blocos), ambos introduzidos por Hayes, Minsky agrega ao termo um terceiro sentido. Os nomes que Schank e Minsky utilizam enfatizam aspectos distintos do que será aqui tomado como uma mesma coisa. Um estereótipo de uma situação inclui, necessariamente, um modo de reconhecer aquela situação como sendo de um dado tipo (se o estereótipo for modelado linguisticamente, então ele conterá critérios para reconhecê-la, por exemplo) Além disso, o estereótipo pode também conter informações sobre como o sistema deve se comportar no interior da situação que ele descreve. O termo "frame" utilizado por Minsky, enfatiza o primeiro aspecto. Por sua vez, o termo "script" utilizado por Schank, enfatiza o segundo. Contudo, tanto um quanto outro permitem os dois tipos de informação. Por isso, ambos podem ser abarcados pelo termo "estereótipo".

os fatos do mundo? Antes, cada fato era marcado como pertencente a um determinado bloco. Da mesma forma, cada ação era marcada como afetando determinado conjunto de blocos. Esses agora precisam ser reclassificados em função do arranjo adequado para a nova situação. A necessidade de reclassificar todos os axiomas presentes no sistema constitui um desafio análogo ao apresentado pelo aspecto inferencial do FP. Afinal, a depender do número de axiomas presentes, pode ser necessário restringir o conjunto de axiomas reorganizados àqueles que são relevantes na presente situação. Como reconhecer quais são relevantes sem ter que checá-los um por um? Segundo, como o sistema cognitivo reconhece a si mesmo no interior de uma dada situação? Quantas e quais características do ambiente ao seu redor ele deve checar até que possa concluir com segurança que se encontra em uma situação do tipo X e não do tipo Y? De modo mais preciso: quais são as características relevantes na circunscrição de cada um destes cenários?

Esses problemas serão desenvolvidos um pouco mais logo adiante. Antes disso, convém considerar uma segunda abordagem *cheap test*, isto é, uma segunda tentativa de dividir o mundo em blocos a fim de eliminar o FP: *conexões causais*. A expressão "causal" não deve ser aqui compreendida estritamente no seu sentido usual. Trata-se de uma relação no interior de um sistema formal que busca modelar relações causais do mundo real. Dizer que A e B são causalmente relacionados significa dizer que as ações que alteram uma propriedade em A podem também alterar esta propriedade em B. Assim, se puder ser provado que A e B não são causalmente relacionados, segue-se que a propriedade de B permanece inalterada. A estratégia é classificar o mundo não por meio das propriedades dos objetos, mas sim a partir da proximidade física entre estes objetos. Quanto mais próximo do evento desencadeado pela ação, maiores as chances de um efeito propagado afetar o objeto. Assim, os efeitos das ações não são categorizados a partir do tipo de propriedade que afetam, mas sim do quão próximos os objetos se encontram do epicentro da ação. Em síntese, o universo é encarado como um grande jogo de pega-varetas.

Uma primeira dificuldade surge ao notar que relações deste tipo não são discretas, mas contínuas. É preciso especificar um critério que permita dizer não apenas quando dois objetos estão causalmente relacionados, mas também em que medida. O critério precisa acomodar gradações, afinal, por vezes um contato indireto entre dois objetos permite propagação dos efeitos (os dois podem estar suficientemente próximos sobre uma mesa, por exemplo). Isto pode levar à necessidade de formular alguma espécie de coeficiente de proximidade causal, que precisa ser suficientemente flexível a fim de lidar com os diferentes casos. No entanto, é duvidoso que tal coeficiente possa ser elaborado. As relações causais entre os objetos são dinâmicas e variam conforme as circunstâncias específicas (o arranjo das varetas na ocasião). Isso significa que a alteração das relações causais entre os objetos é um dos possíveis efeitos de qualquer ação, o que gera a necessidade de mapeá-los. Porém, este mapeamento parece inviável porque pode variar ainda aquilo em função de que se dá uma relação

causal: nem todas as propriedades dos objetos serão afetadas em todas as circunstâncias. Uma vareta A pode ter uma relação causal com a vareta B em função de estar posicionada sobre ela. Assim, a movimentação de A pode gerar a movimentação de B, mas não apenas isso: em certas circunstâncias, aproximar A de B pode fazer com que a temperatura de B seja também modificada. Deve a temperatura ser checada em toda circunstância? Se não todas, em quais? De modo mais geral: quando é preciso checar a ocorrência de alterações naquilo em função de que uma relação causal se deu?

Posto desta forma, isto traz à tona o problema visto há pouco: quão refinados devem ser os mapeamentos das relações causais? Um critério muito permissivo pode incluir um grande número de objetos a considerar, deixando em aberto uma grande quantidade de possibilidades a checar em cada caso, indo na contramão da ideia de realizar poucos testes. Um critério muito específico pode, por sua vez, reduzir o número de objetos a considerar, mas fazer com que o sistema ignore efeitos relevantes recorrentemente. Assim, existe aqui também uma pressão para que haja uma tipificação de segunda ordem abarcando os cenários e as circunstâncias em que cada relação causal se mostra. Retornam assim o problema de como reorganizar as relações causais em função das circunstâncias e o problema de como detectar a situação em que o sistema cognitivo atualmente se encontra. As duas abordagens apresentadas parecem levar à conclusão de que é preciso tipificar os cenários e as circunstâncias em que o sistema cognitivo pode se encontrar. A aposta é que isto permita lidar com universos mais complexos que os de DR31FUS, na expectativa de que viabilizem, por fim, lidar com o universo real.

No entanto, a estratégia estereotípica, quando usada para modelar a LISC tal como esta se manifesta no comportamento humano, sofreu críticas severas de autores como Dreyfus (1987). Se suas críticas são pertinentes, os problemas apontados podem ser fatais à esta abordagem.[18] Para Dreyfus, a própria ideia de que estereótipos pré-definidos podem capturar o comportamento humano na sua totalidade e complexidade, sem que o formalismo torne-se rapidamente não computável, é antes um ato de fé do que um apelo às evidências. Estereótipos são excessivamente descontínuos, isto é, não capturam as nuanças de cada situação concreta. Disso resulta um problema cuja dificuldade é análoga ao de elaborar um número gigantesco de axiomas de *frame*: a necessidade de elaborar estereótipos para todas as variações de cenário concebíveis no âmbito do comportamento humano. Não há uma única maneira de ir a um restaurante ou de se portar em uma sala de aula. Uma vez no restaurante, o

[18] Vale notar que estas críticas não tem como alvo direto as abordagens que buscam tratar da inteligência geral, como era o caso do projeto de Hayes. Porém, elas caracterizam problemas para quem busca modelar o aparato cognitivo humano, tarefa esta que posteriormente se desprendeu da GOFAI, vindo a constituir ramos da psicologia cognitiva. Isso significa que as críticas aqui apresentadas trarão consigo alguns elementos já pertinentes à caracterização do FP enquanto problema filosófico (a ser ainda devidamente apresentado) para muito além de questões envolvendo o ferramental técnico e conceitual utilizado no âmbito da GOFAI.

indivíduo pode, por exemplo, esperar uma mesa; esperar uma mesa acompanhado de alguém; esperar uma mesa acompanhado de alguém no restaurante que é o seu favorito; esperar uma mesa acompanhado de alguém no restaurante que é o seu favorito sentindo-se frustrado porque seu prato predileto está em falta, e assim por diante. A abordagem parece logo reduzir-se a uma tentativa enciclopédica de predizer todos os possíveis cenários e suas possíveis variações.

Contudo, é possível insistir que, de algum modo, qualquer sistema cognitivo, artificial ou não, precisa reconhecer cenários familiares. Embora seja pouco plausível supor que uma enciclopédia de estereótipos tão descomunal venha a constituir uma solução, não é certo que essa crítica constitua argumento suficiente para eliminar por princípio a possibilidade de que a solução seja alguma forma de estereotipação. Mostra-se aí a importância do segundo ponto levantado por Dreyfus: como saltar de um estereótipo ao outro?

> . . . suponha que no estereótipo "vendo alguém que você conhece" a conversa com esta pessoa lhe dê novas informações. Seu amigo pode lhe dizer que um velho amigo estará na cidade somente por mais uma hora, ou que ele acabou de ver seu filho correndo pela rua desacompanhado, ou que o trabalho está lhe deixando maluco, ou que uma empresa cujas ações você possui está prestes a ser vendida, ou que alguém que você vem tentando evitar está comendo na sala ao lado, e assim por diante. Cada novidade pode lhe colocar em uma direção diferente. Do ponto de vista dos estereótipos, o próximo estereótipo poderia ser qualquer um destes: concluir a refeição o mais rápido possível, sair correndo do restaurante, pedir somente um prato rápido, telefonar para seu corretor ou pedir por uma mesa no terraço. Para dar a um computador a capacidade de lidar com esse tipo de mudança, a abordagem dos estereótipos precisaria incluir regras sobre como selecionar o próximo estereótipo. Parece inacreditável que alguém possa escrever, ou efetivamente usar tais regras, e de fato ninguém sequer tentou descrevê-las. (1987, p. 107)

A importância deste ponto pode ser enfatizada por uma rápida revisão das dificuldades enfrentadas pela abordagem *cheap test*. Como se viu, axiomas de *frame* parecem ser válidos em algumas circunstâncias e não em outras. Para mapear estes circunstâncias, dividiu-se o mundo em blocos por meio de critérios como tipos de propriedades ou relações causais. Blocos, porém, também se mostraram sensíveis às circunstâncias, trazendo à tona a necessidade de modelá-las por meio de estereótipos. O ponto de Dreyfus é mostrar que a seleção de um estereótipo como adequado para lidar com as circunstâncias presentes é, ele mesmo, um processo dependente das circunstâncias. De outro modo: se os estereótipos guiam os processos cognitivos, e se a seleção de estereótipos é, ela mesma, um processo cognitivo, como ele se dá? Ainda que se tente formular algum tipo de "estereótipo de segunda ordem", o problema se manifestaria novamente, desta vez na dificuldade em estabelecer um critério de escolha para estes estereótipos de segunda ordem. No fim das contas, não parece ter havido progresso significativo na busca por uma solução ao FP, pois permanece intocado o problema de como encontrar os elementos relevantes em cada circunstância.

Continua necessário explicar no que se sustenta a tartaruga em que se sustentam os elefantes que suportam a terra.

Haugeland apresenta uma ilustração útil na compreensão desta dificuldade com estereótipos. Ele compara o que seriam estereótipos de salas de estar com o estereótipos de cavalos (1987, p. 85–86). Cavalos podem ser descritos em função da semelhança física que apresentam entre si. Eles não apenas tem patas, mas tem quatro patas, duas orelhas e são feitos de carne e ossos em um arranjo que apresenta pouca variação. Disto se segue, por exemplo, que há relativamente poucos modos de cavalgar, o que facilita o mapeamento dos efeitos e não efeitos envolvidos nesta atividade. Salas de estar, por outro lado, não parecem descritíveis dessa forma: tamanho, quantidade de janelas ou arranjo dos móveis podem variar enormemente. Além disso, há uma infinidade de usos para elas: festas, reuniões, jantares etc. Cada um destes modos de usar uma sala de estar apresenta um número igualmente grande (e potencialmente infinito) de variações. Assim, a tentativa de elaborar o estereótipo de uma sala de estar pode se mostrar ou excessivamente detalhado e restritivo em sua aplicação, ou excessivamente vago.

Uma formulação detalhada pode apresentar um grau de especificidade condizente com um estereótipo como o do cavalo, mas isto impede que o modelo seja usado com a flexibilidade com que salas de estar são utilizadas no mundo real. Nas situações não previstas no estereótipo, o sistema cognitivo permaneceria sem saber distinguir o que é relevante do que não é. E com certeza haveria muitas delas, visto que um maior detalhamento implica menor número de situações concretas a que o estereótipo é aplicável. Cada informação adicional funciona como um critério de especiação, reduzindo o escopo de situações concretas em que o estereótipo resultante pode atuar como guia. Assim, o estereótipo acaba funcionando como uma medida *ad hoc* por parte do *designer* do sistema cognitivo ou como uma ferramenta inadequada para quem quer modelar o aparato cognitivo humano, pois ele apenas determina o comportamento adequado para um dado conjunto restrito de situações.

Como seria então o caso oposto? Um estereótipo muito vago, com poucas especificações, também não constituiria uma solução adequada. A vagueza implica que muitas propriedades de uma sala de estar concreta permanecem não mapeadas. Assim, o conjunto de efeitos e não efeitos das atividades ali realizadas precisarão ser determinados no momento da aplicação do estereótipo, e isto deve se dar de modo sensível às circunstâncias. Isto significa, contudo, que os efeitos e não efeitos de ações envolvendo aquele estereótipo podem precisar ser calculados caso a caso a partir de aspectos da situação que não estão presentes no estereótipo. Ora, esse é precisamente o problema para o qual o uso de estereótipos seria uma solução: guiar o sistema cognitivo nos processos, de modo que ele consiga ater-se aos potenciais efeitos e não efeitos relevantes nas circunstâncias em que se encontra. Se o estereótipo apenas subdetermina os efeitos e não efeitos de ações naquelas circunstâncias, ele não constitui solução adequada.

1.3.2 Estratégias *sleeping dog*

Haugeland distinguiu um segundo tipo de estratégias que convencionou chamar de *sleeping dog*.[19] Esse tipo de estratégia opera sob um prisma diferente. Em vez de repartir o universo em blocos de modo a restringir a especificação das regras e a realização dos cálculos a esses blocos específicos, busca-se implementar a LISC de maneira mais direta: *"... deixe tudo como está, a menos que tenha alguma razão positiva para não fazê-lo."* (HAUGELAND, 1987, p. 84). Assim, ao invés de tomar a LISC como um conjunto de regras, busca-se modelá-la como um modo *default* de raciocínio. No caso de DR31FUS, por exemplo, a ideia seria utilizar uma cláusula que permitisse ao sistema supor que, na ausência de indicações explícitas em contrário, tudo o que era verdadeiro em S0 permanece verdadeiro em S6. O desafio, claro, é encontrar um critério adequado para lidar com os casos em que nem tudo permanece verdadeiro. O que conta e o que não conta como uma razão para deixar de ignorar um possível efeito de uma ação?

Um primeiro exemplo deste tipo de abordagem é a estratégia utilizada pelo STRIPS (*Stanford Research Institute Problem Solver*) de NILSSON (1971). Tal como no caso do cálculo de situação, um cenário é representado por um conjunto de axiomas expressos como fórmulas análogas às de lógica de primeira ordem. A fórmula abaixo, adaptada a partir de NILSSON (1971), exemplifica como a situação S0 de DR31FUS seria descrita:

$$S0 : atr(a) \wedge at(MCD, m) \wedge at(TOD, m) \wedge at(MPT, TOD) \wedge at(HEI, MPT)$$

A operação `atr(a)` indica que o robô está na posição a (onde a pode ser um vetor n-dimensional que inclui, por exemplo, um valor para profundidade, outro para posição vertical e outro para posição horizontal). Já o termo m indica a posição da mesa. Assim, `at(MCD,m)` pode ser lido como: o bloco MCD está sobre a mesa. Também tal como no cálculo de situação, os objetivos do sistema cognitivo (isto é, o "problema" a resolver) são expressos por meio de fórmulas como:

$$G1 : at(TOD, MCD) \wedge at(MPT, TOD) \wedge at(HEI, MPT)$$

A fórmula G1 acima expressa uma situação alvo em que o bloco TOD esteja sobre o bloco MCD, o bloco MPT esteja sobre o bloco TOD e o bloco HEI esteja sobre o bloco MCD. Essa é precisamente a situação obtida em S6 após a realização das operações necessárias (vide fig. 1). Até o momento, contudo, nenhuma diferença substancial em relação ao cálculo de situação fora apresentada. A distinção do STRIPS começa a se mostrar no modo como são modeladas as ações que o sistema cognitivo

[19] O nome vem do dístico *"let sleeping dogs lie"*. Esta expressão é usada quando se quer recomendar a alguém que deixe alguma situação presente no estado em que se encontra, pois mexer com ela pode fazer com que as coisas piorem.

pode realizar, denominadas *operações*. A especificação se dá por meio de três listas. Cada uma destas listas contém fórmulas análogas às utilizadas para descrever situações e os objetivos do sistema cognitivo. A primeira estabelece os pré-requisitos, isto é, os axiomas que precisam ser verdadeiros para que a operação seja possível. As outras duas listas estabelecem os efeitos da ação: uma especifica os axiomas a ser adicionados ao conjunto de fórmulas que descreve a situação e a outra especifica os axiomas a remover. De outro modo: as listas especificam os fatos que precisam ser verdadeiros para permitir a realização da ação, os fatos que passam a ser verdadeiros uma vez que a ação tenha sido realizada e os fatos que deixam de ser verdadeiros.

Em DR31FUS, as listas que compõem uma operação como `mover_objeto(k,m,n)` podem ser especificadas facilmente.[20] Esta operação pode ser interpretada como uma ordem para o braço mecânico: movimente o objeto k, do local m até o local n. Uma versão bastante simplificada da lista de pré-requisitos poderia ser:

$$REQ : atr(m) \wedge at(k, m)$$

A semântica deste requisito é: tanto o braço mecânico quanto o objeto k precisam estar na posição m. A lista de axiomas a adicionar poderia ser:

$$ADD : atr(m); at(k, m)$$

Isso significa: nem o braço mecânico nem o objeto k estão mais na posição m. Por fim, a lista de elementos a remove poderia ter o seguinte conteúdo:

$$DEL : atr(n); at(k, n)$$

O significado: o braço mecânico e o objeto k estão agora na posição n. Assim, o conhecimento do sistema cognitivo sobre o que costuma caracterizar ou não os efeitos de uma determinada ação é modelado nas listas que constituem as operações do sistema.

Essa forma de modelar ações ainda não traz nenhuma diferença significativa da estratégia utilizada no cálculo de situação. Afinal, o FP poderia ser descrito como a necessidade de uma quarta lista contendo os axiomas que *não* serão afetados pela ação. Contudo, o que caracteriza o STRIPS como uma estratégia *sleeping dog* é precisamente a ausência desta lista. A LISC não é modelada diretamente, mas é tomada como um modo *default* de operação com base em algumas táticas. A primeira delas: a mitigação da abordagem dedutivista. STRIPS não abandona métodos de prova de teoremas tais como os usados no cálculo de situação, mas restringe seu escopo: estes são usados apenas para responder perguntas acerca do atual estado de coisas. Tais perguntas podem ser feitas, por exemplo, ao checar se os pré-requisitos de uma ação estão satisfeitos ou ao verificar se um determinado objetivo como G1 fora satisfeito. Em outras tarefas, tais como determinar o melhor curso de ação a realizar ou a

[20] Exemplo adaptado de NILSSON (1971) p. 197.

relevância de um elemento, o método aplicado é análogo ao do GPS. Fikes sintetiza a abordagem da seguinte forma:

> Métodos de prova de teoremas são usados apenas dentro de um dado modelo de mundo para responder questões acerca de quais operadores são aplicáveis a ele e se algum dado objetivo foi ou não satisfeito. Para procurar através do espaço de modelos de mundos, STRIPS faz uso de uma estratégia de análise "meios-fins" nos moldes do GPS. (NILSSON, 1971, p. 190)

O espaço de modelos de mundos referido por Fikes é o conjunto de situações que podem vir a ser o caso. O sistema não se pauta por uma modelagem prévia das circunstâncias que ditam a situação que virá a ser o caso, mas por operações falíveis envolvendo uso de heurísticas, tentativa e erro e táticas análogas. Estas operações tentam fazer "palpites bem informados" sobre quais são as fórmulas com mais chances de minimizar o número de inferências necessárias para que se possa chegar de uma situação S0 à S1. Tais heurísticas podem, por exemplo, testar diferentes cenários e avaliar em qual deles o sistema tem mais chances de atingir o objetivo com menos cálculos. Desse modo, o sistema pode partir do princípio de que todo efeito não explicitamente catalogado nas operações pode ser desconsiderado. Eventuais exceções teriam seu espaço na medida em que surgissem durante a busca não dedutivista por soluções possíveis aos problemas. O resultado é uma estratégia falível, mas em geral confiável. [21] A assunção de que todo efeito ou não efeito que não fora explicitado pode ser ignorado, característica que distingue a abordagem *sleeping dog*, evita casos como aquele em que DR31FUS não conseguia responder se um bloco movimentado permanecia vermelho. O uso de modos não dedutivos de inferência, por sua vez, permite ao sistema detectar com razoável confiabilidade situações excepcionais, como aquelas que envolvem movimento de objetos em cenários com tinta fresca ou camaleões. Em síntese, para Fikes e Nilsson, o FP era fruto do dedutivismo. Ao abrir mão do cálculo de provas de teoremas para vasculhar possíveis resoluções de problemas, sentiram-se justificados em afirmar que o FP havia sido superado.

Fosse o FP incapaz de resistir a isto, certamente não estaria sendo investigado ainda hoje. STRIPS apresenta uma estratégia que é, *prima facie* razoável para lidar com o FP. O uso de heurística pode vir a abrir o caminho para uma solução algorítmica do tipo que este aspecto demanda. Contudo, o FP retorna porque o STRIPS não evita a necessidade de decisões caracteristicamente *ad hoc* na hora de modelar as operações do sistema. Isto se mostra da seguinte forma: as listas de efeitos (fórmulas a adicionar e a remover) de uma operação qualquer não podem ser específicas demais. Elas não podem conter absolutamente todos os efeitos possíveis. O motivo: tamanha especificidade impede seu uso em pouco mais que alguns casos concretos em cená-

[21] O STRIPS antecipa, em larga medida, o que será posteriormente apresentado como o antagonismo entre a posição de Fodor, McDermott e Samuels. Isto sugere que boa parte do debate filosófico em torno do FP é uma versão menos ingênua de debates que se deram no âmbito da GOFAI. Esta é uma das principais razões pelas quais se dedica tal espaço a essas análises.

rios muito específicos e artificialmente estabelecidos (tais como micro-universos). Em um universo complexo, é preciso reservar algum espaço de manobra para que os efeitos concretos variem em função das circunstâncias específicas em que a operação se dá, tal como é característico do comportamento inteligente. A insistência em descrições detalhadas faz com que STRIPS recaia na necessidade de postular algum outro tipo de estratégia para lidar com os diferentes cenários em que as operações podem ser realizadas. Logo, o sistema demanda algum tipo de estereótipo a partir do qual as operações podem ser definidas, tal como exposto no caso das abordagens *cheap test*. Junto disso, retornam também os problemas relacionados ao uso de estereótipos.

Para resistir ao uso de estereótipos, STRIPS usa uma tática alternativa: a divisão dos fatos sujeitos à mudança entre fatos *primitivos* e fatos *não primitivos*. O objetivo é fazer com que todo efeito não explicitamente mapeado dê-se sempre como um fato não primitivo. Fatos primitivos são fatos que não possuem qualquer interdependência entre si. A posição de uma peça num jogo de xadrez é um exemplo de fato primitivo, visto que, dentro do conjunto de movimentos permitidos pelas regras do jogo, movimentar uma peça não afeta nenhum outro fato primitivo.[22] Assim, quando uma ação é realizada, o sistema não precisa checar possíveis efeitos sobre nenhum outro fato primitivo, podendo assumir que todos permanecem inalterados de uma situação à outra. O uso de heurística se restringe à busca por fatos não primitivos que podem vir a ser relevantes. Esses fatos não primitivos são sempre deriváveis de fatos primitivos, e não precisam constar explicitamente no conjunto de axiomas que descrevem o cenário, visto que podem ser derivados “sob demanda” conforme as circunstâncias.

Uma primeira consequência desta estratégia é um desafio que o STRIPS deve superar se quiser fornecer uma solução para o FP. Boa parte da sua viabilidade se deve à escolhas técnicas sobre quantos dos fatos não primitivos serão armazenados explicitamente na memória e quanto deles serão deixados para derivar sob demanda. É possível restringir-se a armazenar somente os fatos primitivos, deixando os fatos não primitivos sempre implícitos. Desse modo, a alteração de um fato primitivo irá afetar tudo o que se pode derivar do axioma que o representa sem que seja necessária qualquer ação adicional. Essa tática, no entanto, coloca um peso excessivo sobre a necessidade de processamento, que é um bem mais valioso do que a memória. Como explicitado anteriormente, enquanto a memória pode até ser tomada como tendo limites desprezíveis, o processamento está ligado ao tempo e sofre das limitações físicas a ele relacionadas. Há um limite para o quão rápido um evento físico pode ser e há um limite para o quão razoável é esperar que um processo seja concluído. Por isso, em versões menos radicais, alguns fatos não primitivos são também explicitamente armazenados.

No entanto, ambas as táticas precisam enfrentar uma mesma dificuldade: quando parar de derivar? O problema pode ser visto tanto do ponto de vista do sistema (em

[22] Evidentemente, um movimento pode fazer com que uma outra peça seja retirada do jogo, mas isto é parte do que constitui uma ação previamente mapeada, e não um efeito possível a ser checado.

que momento ele deve parar de derivar fatos não primitivos a partir de fatos primitivos?) quanto do ponto de vista do *designer* do sistema (como especificar um critério ou heurística não *ad hoc* para que o sistema possa decidir quando parar?). A opção pelo armazenamento de fatos não primitivos não elimina esta dificuldade, mas antes faz com que ela se manifeste de outro modo. A cada evento que altere um fato primitivo, o sistema deverá revalidar todos os fatos não primitivos que dele derivam. Porém, fatos não primitivos podem apresentar interdependências entre si, sendo necessário avaliá-las também, trazendo novamente a pergunta: quando parar? Uma caracterização nestes termos foi o que fez com que Fodor descrevesse o FP como sendo *"... o problema de Hamlet do ponto de vista de um engenheiro"* (1987, p. 140), isto é, quando parar de pensar?

A maior dificuldade, no entanto, não é esta. Como visto, a distinção entre fatos primitivos e fatos não primitivos é o que permite afirmar que STRIPS faz uso da LISC sem modelá-la na forma de um conjunto de axiomas de *frame*, blocos ou estereótipos. Há uma outra dificuldade, talvez fatal, apontada por Haugeland (1989): como determinar o que será tomado como um fato primitivo? Uma tal distinção parece separar o universo entre o que é seu esqueleto, seu pilar de sustentação e o que caracteriza o conteúdo a ele agregado. Categorizar fatos desta forma é análogo a prover uma descrição de caráter ontológico de um universo. Tarefa hercúlea, quando o objetivo é gerar um modelo fiel ao universo real, e não um micro-universo. Não bastasse, ressurge a dificuldade de sempre: os critérios que justificam tal categorização de fatos parecem variar conforme a circunstância. Se a posição de uma xícara é tomada como um dado primitivo, então mover outro objeto no universo considerado nunca deveria afetá-la, mas se ela estiver sobre um pires, movê-lo irá afetá-la. Ainda que se tente algo mais elaborado (tomar como um fato primitivo a posição relativa de um objeto a aquilo que o sustenta, por exemplo), não há como supor que isto funcionará em todos os cenários, ou sequer que será suficiente para funcionar nos cenários relevantes: a depender de como se movimenta o pires, a xícara pode cair ou ser afetada de outro modo, dificultando que sua posição relativa seja tomada como básica.

O cenário apresentado traz à tona a pergunta: o que poderia contar como um motivo suficientemente forte para reconhecer as circunstâncias atuais como circunstâncias que demandam uma revisão no que é considerado um fato primitivo e no que não é? Tal como no caso da estratégia *cheap test*, o resultado é uma demanda por algum tipo de categorização das circunstâncias em que as operações são realizadas, tais como as que se tenta fazer via estereótipos. Assim, a própria organização das listas que constituem as operações parece necessitar de revisão em função da situação em que o sistema cognitivo se encontra. Não bastasse a dificuldade de categorizar os fatos do mundo, parece necessário fazê-lo uma infinidade de vezes. No caso das abordagens *cheap test* o resultado final foi o desafio de reorganizar axiomas de *frame*, blocos e/ou conexões causais em função de estereótipos que, eles mesmos, dependiam das circunstâncias para guiar o sistema nesta reorganização. No caso das

abordagens *sleeping dog*, dá-se o mesmo, mas ao invés de reorganizar axiomas ou blocos, é preciso reorganizar os fatos que podem ser considerados primitivos ou não primitivos, a depender das circunstâncias.

Antes de partir para a próxima etapa, convém analisar pelo menos mais uma tentativa concreta de superar o FP por meio de uma estratégia *sleeping dog*. É o caso do *Unless*, elaborado por SANDEWALL (1972). Trata-se de um operador lógico bastante peculiar, formulado especificamente para lidar com o FP, embora ele evidentemente possua outros usos. A ideia de Sandewall é incorporar o *Unless* ao PCF-2, um sistema de cálculo alternativo ao cálculo de situação. Trata-se de um formalismo cuja sintaxe é muito semelhante à lógica de primeira ordem, mas cujos operadores podem exibir propriedades distintas. Além disso, o PCF-2 possui operadores distintos, como é o caso do próprio *Unless*. Ao contrário do STRIPS, contudo, não há uso de táticas análogas às do GPS. O PCF-2 permanece na órbita das abordagens dedutivistas, portanto. Uma análise do formalismo de Sandewall está fora do escopo desta investigação, mas é possível mostrar, primeiro, por que Sandewall acreditava que o Unless é uma solução ao FP e, segundo, porque ele não é. A forma básica do operador é a seguinte:

$$A \wedge Unless(B) \supset C$$

A propriedade essencial de `Unless()` é considerar que `Unless(B)` está provado em todos os casos em que não for possível provar B. Assim, se for o caso que A e não for o caso que B, segue-se que C. A presença de um operador desta natureza faz do PCF-2 um formalismo que abandona a monotonicidade da lógica clássica. Como sintetiza Shanahan: *"... uma lógica é monotônica se ela garante que a adição de um novo fato jamais poderá fazer com que uma consequência anterior deixe de valer."* (1997, p. 15). Com efeito, em formalismos monotônicos, não há como cancelar uma consequência derivável de um conjunto de axiomas por meio do acréscimo de mais axiomas. Quanto mais axiomas um sistema possuir, mais dele se pode derivar. Assim, implementar um movimento típico como o de revisão de crenças acerca de um cenário em função de novas informações torna-se um desafio. Uma das apostas que fundam o *Unless* é a de que esta propriedade é parcialmente responsável pelo surgimento do FP. A não monotonicidade do PCF-2 é perceptível no seguinte cenário:[23]

$$A, Unless(C) \supset B \vdash B$$

Em um sistema que contenha somente os axiomas acima, B será um teorema, pois pode ser derivado da ausência de C. Contudo, se o axioma C for acrescentado ao sistema, segue-se disto que B deixa de ser um teorema. Assim:

$$A, Unless(C) \supset B, C, \nvdash B$$

[23] Adaptado de JANLERT (1987).

Assim, ao invés de lidar com listas de axiomas a adicionar ou remover como faz o STRIPS, o PCF-2 faz uso de descrições de ações muito mais próximas daquelas utilizadas no cálculo de situação (axiomas de efeito). A diferença, claro, é a disponibilidade do operador `Unless()` para a definição de tais axiomas. Consegue-se então um meio termo entre formular poucas regras completamente insensíveis às circunstâncias (a cor do objeto não muda nunca) e a necessidade de formular regras para cada possível ação no universo em que a propriedade se insere (mudar o bloco de lugar não afeta a cor, fazer uma conta de matemática não afeta a cor, espirrar não afeta a cor etc.) O operador permite formular regras para ações, propriedades e objetos sem a necessidade de formular axiomas de *frame* para os não efeitos de cada uma delas. As regras equivalentes aos axiomas de efeito podem ser formulados de modo a especificar as ocasiões em que uma dada propriedade deve ser verificada. Algo como: "se puder ser provado que o bloco é vermelho, e não puder ser provado que o bloco deixou de ser vermelho, então o bloco permanece vermelho". Dessa forma, não é preciso formular uma regra (um axioma de *frame*) que explicite a manutenção da cor de um bloco para todas as possíveis ações. Se as condições especificadas a partir do operador `Unless()` não forem satisfeitas, a propriedade mantém seu valor atual através das situações, razão pela qual esta abordagem se insere no conjunto *sleeping dog*. Desse modo, a expectativa é que o uso desse operador permita escrever axiomas de efeito suficientemente flexíveis para que as ações possam ser confiavelmente aplicadas em situações concretas.

A ideia de que um operador como o `Unless()` pode, por si só, evitar o FP, é sujeita a diversos questionamentos. O primeiro deles liga-se à viabilidade de implementá-lo. Em cenários suficientemente simples, a conclusão de que um dado teorema não pode ser provado é relativamente simples de se atingir. Contudo, na medida em que o cenário se complexifica, a tarefa se torna cada vez mais difícil, a ponto de gerar uma versão local do FP, a esta altura já familiar: quando parar de checar? Isso é, se o sistema cognitivo encontrar-se em um ambiente suficientemente complexo de modo que não seja possível checar todas as alternativas que lhe permitiriam concluir que um dado teorema não pode ser provado, como saber quantas devem ser checadas antes de parar? Quando terá o sistema checado possibilidades "suficientes" para justificar a parada? Esta é uma dificuldade de natureza idêntica à que emergiu no caso do STRIPS. Lá havia o problema de decidir um critério para quando o sistema cognitivo deveria parar de derivar fatos não primitivos. Aqui o problema se mostra como a dificuldade em definir um critério de parada para casos em que não é possível concluir de modo eficiente que um dado teorema não pode ser provado a partir de um dado conjunto de axiomas.

Ainda que este problema possa ser mitigado ou mesmo resolvido, também como no caso do STRIPS, a dificuldade maior é outra. Conforme JANLERT (1987) aponta, o *Unless* não evita o problema da especificidade com que as ações são descritas. Ele oferece a seguinte ilustração: suponha uma janela composta por vários peque-

nos vidros separados em uma estrutura e uma ação como `jogar_pedra_na_janela()`. A pedra é jogada e quebra um dos vidros, mas a ação não consegue fornecer as pistas para provar qual deles foi quebrado. Na impossibilidade de provar que um dos vidros está quebrado, o sistema conclui que nenhum deles se quebrou e que a janela permanece intacta. Isso mostra que o uso do `Unless()` é, ele mesmo, dependente das circunstâncias: nem sempre a impossibilidade de provar x equivale à ausência de efeitos relevantes envolvendo x.

Essa ilustração dá margem a questionamentos como: não pode esta ser uma limitação perceptual ao invés de uma limitação na caracterização da inteligência? Mesmo um ser humano precisaria olhar para a janela se quisesse concluir qual dos vidros quebrou ou, ao menos, ouvir o barulho para concluir que algum dos vidros quebrou. Porém, a questão desvia do ponto principal: não pode ser o caso que todos os cálculos de efeitos e não efeitos sejam resolvidos a partir de *inputs* perceptuais. Deve haver algum modo "interno" de lidar com este problema de especificidade. Mesmo de olhos fechados, deve ser possível concluir que mover um bloco não altera sua cor ou que mover um pires leva consigo a xícara que repousa sobre ele, quando é o caso. Trata-se de um problema muito similar ao descrito no caso do STRIPS: ações cujos efeitos tenham sido ricamente mapeados acabam sendo pouco flexíveis e servindo apenas para um pequeno conjunto de casos. Ações cujos efeitos sejam mapeados de forma vaga, contudo, acabam deixando o sistema "cego" para um grande conjunto de efeitos. O *Unless* termina por deixar o *designer* de sistemas cognitivos na mesma situação em que o STRIPS lhe deixara.

O que se segue destas considerações? Antes de partir para caracterização filosófica do FP, cabe uma rápida revisão das conclusões preliminares obtidas. Para abordagens *cheap test*, a LISC deve ser modelada como um conjunto de critérios que permitem selecionar partes do universo. Cada parte do universo circunscreve os elementos que devem ser considerados relevantes na situação. Ao categorizar, crê-se ser possível encontrar regras gerais que possam reger as relações entre os tipos determinados, reduzindo assim o número de operações computacionais a um nível tratável. O FP seria resolvido pela redução drástica da quantidade de possibilidades que o sistema cognitivo precisa checar em cada situação concreta. Assim, aquilo que soa uma obviedade (concluir que um bloco movido não altera sua cor), pode ser tratado de forma eficiente. Porém, essa tipificação não parece dar conta do modo como as situações concretas se estruturam. Dividir o mundo em blocos mostra que essa divisão é sensível às circunstâncias. Inserir os blocos em situações estereotipadas (bares, restaurantes etc.) para lidar com essa dependência mostra que os próprios estereótipos sub-determinam as circunstâncias. Além disso, há o desafio de explicar como o sistema decide qual deve ser o estereótipo adequado. Em última instância, o sistema ainda não consegue circunscrever os efeitos e não efeitos relevantes nas circunstâncias presentes, ou seja, não consegue fazer uso da LISC.

No caso das abordagens *sleeping dog*, a LISC deve ser modelada não como

uma cláusula explícita do sistema, mas como um modo de realizar operações. Assim, não é preciso testar se é o caso que o movimento de um bloco afeta sua cor ou quaisquer outras propriedades, a menos que haja um bom motivo para considerar esta possibilidade. O problema reside em determinar o que pode ser considerado um bom motivo. Algum teste para checar se é o caso deve, eventualmente, ser feito. Quando? A tática do STRIPS é dividir o universo entre seus pilares de sustentação (os fatos primitivos) e o restante do prédio construído sobre estes pilares (os fatos não primitivos). Assim, quando um fato primitivo é afetado por uma ação, checa-se por possíveis efeitos em fatos não primitivos que dele derivem. Se, por outro lado, uma ação não afeta um determinado fato primitivo, ele é deixado no estado em que se encontra e nenhum dos fatos não primitivos que dele derivam é checado. No entanto, essa tática apresenta o mesmo tipo de insensibilidade às circunstâncias que a tentativa de dividir o mundo em blocos possui, afinal, a própria categorização entre o que é um fato primitivo e o que não é varia circunstancialmente. No caso do operador *Unless*, a tática é considerar que a impossibilidade de provar um teorema implica sua negação. Assim, aquilo que conta como um bom motivo é o que está especificado por meio de cláusulas com o operador `Unless()`, permanecendo todo o resto inalterado.

Contudo, assim como no STRIPS, a tensão entre a necessidade de descrições de ações que sejam, ao mesmo tempo, suficientemente detalhadas (sem vagueza), e suficientemente flexíveis (com vagueza), traz à tona o problema de que mesmo o uso do operador `Unless()` é dependente das circunstâncias. Os dois casos parecem demandar algum tipo de tratamento mais acurado das circunstâncias. A classificação dessas por meio de estereótipos surge no horizonte. O problema: se as críticas de Dreyfus forem corretas, utilizar estereótipos não é uma opção. Como se viu, todas as tentativas de dividir ou classificar aspectos do universo de modo *a priori* resultaram ou em quantidade excessiva de falhas na captura de cenários relevantes, ou em quantidade excessiva de critérios a determinar e cálculos a realizar. A tartaruga permanece no espaço, sem que se consiga explicar o motivo de ela não cair. Fossem os estereótipos um passo na direção certa, o FP não seria um desafio tão monumental, mas sim uma mera dificuldade de implementação. Isso sugere que a inteligência humana não se pauta (apenas) por tipos pré-definidos, sejam estes naturais, sejam convencionais. Esta é uma ideia que caracterizará boa parte do debate acerca do FP no âmbito da filosofia.

1.4 De problema técnico a problema filosófico

Até o momento, o FP foi predominantemente tratado como um problema de caráter técnico e ligado às dificuldades de implementação características dos formalismos utilizados no âmbito da GOFAI. Contudo, o fato de que o FP tenha se mostrado não apenas em tentativas de desenhar quaisquer sistemas que busquem implementar

uma inteligência ideal, mas também na tentativa de modelar componentes da mente humana, permitiu que o problema fosse tomado como extrapolando os objetivos mais específicos da GOFAI. Para muitos autores, o FP não é um problema apenas para quem quer modelar a mente humana fazendo uso das ferramentas disponibilizadas no interior da GOFAI, mas sim um problema mais amplo, conceitual, e que vai ao cerne do movimento cognitivista. Essa transição de problema técnico a problema filosófico não foi suave.

Alguns anos após sua primeira descrição por Hayes, Daniel Dennett (1999[1978]) sugeriu que o FP teria relevância filosófica. Ele argumentava pela importância do papel que a IA poderia exercer na solução de problemas filosóficos clássicos e no desvelamento de novos problemas. O FP seria um exemplo de problema epistemológico que se manifestou no âmbito das pesquisas da GOFAI: como manter um conjunto de crenças sobre o mundo devidamente atualizado de modo consistente? Mais especificamente, como explicar a capacidade humana de manter um tal conjunto de crenças atualizado? Nessa concepção, o FP não se restringe a uma dificuldade de implementação de uma versão artificial da inteligência, mas refere-se a uma questão importante acerca da possibilidade de explicá-la.

> Não é preciso ter a esperança de um futuro cheio de robôs para que o *frame problem* seja uma preocupação. O problema aparentemente emerge de pressuposições largamente aceitas e aparentemente inócuas sobre a natureza da inteligência, a verdade da vertente menos doutrinária do fisicalismo e a convicção de que deve ser possível explicar como pensamos. (1987, p. 44)

Posteriormente, Dennett (1987) defende a ideia de que o FP técnico, tal como se manifestou na IA, ainda que não seja pouco desafiador ou importante por si só, era apenas uma versão desse problema epistemológico mais geral. Para ele, é este problema epistemológico, e não apenas a manifestação dele enquanto problema técnico numa empreitada específica, que merece a alcunha de FP. Perspectivas como essa, somadas à dificuldade de compreensão do problema técnico original em sua inteireza deram origem a um acalorado debate no decorrer dos anos 1980-90 acerca da natureza do FP. O debate abrangeu muitas divergências sobre a relação entre as considerações filosóficas do FP e sua versão técnica. Essas divergências resistem, em certa medida, até hoje. Por isso, antes de firmar compromisso com alguma concepção filosófica do FP, é preciso assumir uma posição na discussão de fundo acerca da própria pertinência do FP no âmbito da filosofia.

1.4.1 O que está em jogo no debate filosófico

Para facilitar a apresentação do debate, a expressão “FPIA” será utilizada deste ponto em diante para referir-se ao *Frame Problem* enquanto problema técnico e lógico, nos moldes previamente expostos. A expressão “FP”, por sua vez, será utilizada para referir-se apenas ao *Frame Problem* enquanto problema filosófico, cuja definição

permanece em aberto. Diante dessa distinção, surge uma primeira pergunta: qual a relação entre FP e FPIA? São esses problemas distintos ou diferentes versões de um mesmo problema? Se são problemas distintos, qual a natureza desta distinção? Haverá algum grau de parentesco, ainda que distante, ou a relação entre eles é fruto de mera contingência histórica? Para Shanahan (1997), por exemplo, o FPIA pode ser tomado como um problema específico das abordagens dedutivistas que independe de quaisquer considerações filosóficas. Por outro lado, segundo Crockett (1994), a versão técnica do FP não pode ser considerada resolvida até que o problema mais geral subjacente, de caráter filosófico, tenha sido devidamente tratado. Esses posicionamentos envolvem nuanças que ainda não se pode avaliar, mas que virão à tona em breve. Assim, pode-se tomar como ponto de partida a tese de que FP e FPIA são problemas distintos, deixando a natureza exata da relação entre eles para um momento posterior. A tese pode ser sintetizada nos seguintes termos:

Tese 1 (T1): O FP é um problema filosófico não idêntico, embora possivelmente relacionado, ao FPIA.

Dada esta definição, é possível aceitar T1 sem que isto gere compromisso com qualquer caracterização específica do FP enquanto um problema filosófico, bastando que ela seja distinta do FPIA. Pode-se até mesmo negar qualquer relação entre o problema técnico e o problema filosófico, exceto a coincidência do nome e o fato de um ter servido como inspiração para um *insight* que trouxe o outro à tona.

A aceitação ou não de T1 pode variar, portanto, em função do modo como se caracteriza o FP no âmbito da filosofia. Embora Dennett tenha sido o primeiro a fazê-lo, é em Haugeland que pode ser encontrada a caracterização que mais se aproxima de uma reconstrução do FPIA em termos genéricos e desconectados dos seus aspectos técnicos. A descrição é a seguinte: na medida em que os eventos do universo se dão, podem ocorrer alterações no estado de coisas ao redor do agente. Este precisa continuamente *revalidar* suas crenças acerca destes estados de coisas. Isto demanda, primeiro, uma divisão entre fatos *perenes*, ou seja, fatos que podem ser tomados como fixos no contexto, e fatos *temporários*, que estão sujeitos à mudança. Segundo, no caso dos fatos temporários, os critérios de revalidação precisam permitir tanto o mapeamento de possíveis interdependências entre si (como um fato temporário pode afetar o outro) quanto a seleção das interdependências relevantes no contexto específico. A pergunta que surge é: como selecionar as interdependências relevantes? De outro modo: como circunscrever (i.e. determinar um *frame* para) a informação a ser considerada? Note-se que o problema surge mesmo que se restrinja o raciocínio a contextos pequenos e bem definidos: *"... as interconexões e dependências relevantes são elas mesmos dependentes da situação, ou seja, os efeitos colaterais que efetivamente se darão a partir de qualquer evento particular são sensíveis às minúcias das circunstâncias correntes"* (HAUGELAND, 1989, p. 205).

É importante salientar que essa descrição não se restringe nem se relaciona diretamente com a determinação de efeitos ou não efeitos de ações, sugerindo que a origem do problema não tem relação direta com o cenário restrito em que ele primeiro se manifestou. A dificuldade enfrentada ao modelar o aparato cognitivo humano é mais geral que a dificuldade de tratar efeitos ou não efeitos de ações. De algum modo, em sua interação com o universo, a cognição humana parece capaz de ater-se somente aos aspectos relevantes. De modo suficientemente confiável, aquilo que é relevante simplesmente aparece como saliente, sendo pouco plausível supor que, a cada momento, em função de cada percepção, a totalidade do conjunto de crenças seja revalidada. Esta capacidade pode ser demonstrada com um experimento simples: basta fechar os olhos e perguntar a si mesmo qual a cor das letras deste texto. Parece evidente que não é necessário recorrer a *inputs* perceptuais para manter a crença de que a cor das letras permanece a mesma. Ainda que seja possível imaginar casos pitorescos, típicos da reflexão filosófica, em que a cor das letras em um papel mudem por intermédio de um gênio maligno ou de um efeito físico determinado, estas possibilidades geralmente se mostram ao mesmo tempo em que surge um senso de sua implausibilidade ou irrelevância na determinação do comportamento resultante.

Convém enfatizar que a capacidade a ser explicada não é a de acertar sempre, mas sim a de acertar um número razoável de vezes utilizando os recursos cognitivos de modo eficiente. Não seria razoável supor que processos cognitivos busquem a infalibilidade, pois isto implicaria que, em algum nível, eles tenham levado em conta todas essas possibilidades imaginadas, bem como uma infinidade de outras não explicitadas, apenas para então concluir que as cores das letras no papel permanecem as mesmas.[24]

A cognição humana parece dar conta deste problema de modo extremamente eficiente, ainda que falível, visto que tais revalidações se dão em tempo real, na medida em que os diferentes contextos se apresentam. Assim, diante da tarefa de modelar as capacidades cognitivas humanas, este parece um desafio inescapável. Aquilo que surgiu como uma dificuldade técnica na implementação de um sistema de cognição artificial, mostra-se agora como um desafio na elaboração de uma explicação sistemática (isto é, de um modelo) da cognição humana, mas de modo desconectado de quaisquer limitações técnicas inerentes aos formalismos utilizados. Trata-se, portanto, de um desafio para a TCC cuja origem e natureza são ainda desconhecidas. Isso pode ser sintetizado na seguinte tese:

Tese 2 (T2): O FP se mostra em qualquer tentativa de modelar o aparato cognitivo humano.

[24] A abrangência do problema pode ser enfatizada se for notado que aquilo que é levado em conta num processo cognitivo pode envolver, por exemplo, uma história filogenética e uma história ontogenética, que constrangem os caminhos inferenciais a que o aparato cognitivo se mostra disposto a perseguir. Estes aspectos serão considerados, em alguma medida, no terceiro capítulo.

É importante notar que a aceitação de T1 não implica aceitação de T2. Boa parte das discussões sobre o FP tratam das condições de possibilidade para que ele possa manifestar-se. Como visto, na GOFAI, explicar a cognição humana é uma empreitada intimamente ligada à construção de sistemas que realizem as tarefas que a inteligência humana desempenha. Mas é possível resistir a isto. Há vários modos de negar T2, ainda que se aceite T1. Uma possibilidade é negar que a inteligência humana seja plenamente modelável nos termos cognitivistas (isto é, negar ou mitigar o cognitivismo). Neste caso, o FP se mostra como um problema insuperável que demanda não uma solução, mas uma *dissolução*: é preciso fazer uso de alguma abordagem alternativa que permita explicar por que o modo como a mente humana gera comportamento inteligente não permite que o FP venha à tona. Essa é, por exemplo, a posição de DREYFUS; DREYFUS (1987). As características exatas da abordagem alternativa proposta podem variar enormemente de autor para autor, não cabendo analise detalhada aqui. O importante é notar como estas posições se distinguem de aceitar que o problema é, de algum modo, *solucionado* pela mente humana, como se a resposta fosse parte do que constitui o algoritmo da inteligência (exatamente como supõe o cognitivismo). Assim, na empreitada cognitivista, aceitar T2 traz consigo a necessidade de formular uma solução ao FP. Por sua vez, negá-la demanda fornecer algum tipo de alternativa ao cognitivismo, ou uma substancial revisão.

Outra tentativa de aceitar T1 e negar T2 se dá pela defesa de que o FPIA e o FP tem uma origem em comum, mas constituem problemas distintos. O problema que origina a questão teria caráter metafísico e ontológico. Isso demanda alguma elaboração: como visto no caso do STRIPS, a tentativa de categorizar o universo entre fatos primitivos e fatos não primitivos constituía a estratégia utilizada para modelar o universo real. Isso levou a perguntas difíceis como: o que conta como um fato primitivo e o que não conta? O mesmo pode ser dito da tentativa de modelar o universo a partir da divisão do mesmo em blocos, como no caso das abordagens *cheap test*. Quais seriam os blocos adequados?

Para autores como Hayes, este é o verdadeiro desafio por trás do FPIA. Uma vez que se realize o mapeamento das relações primárias entre os elementos constituintes do universo, modelá-los será uma tarefa árdua, mas realizável. Como sintetizou Hayes: *"Não estamos recortando a natureza nas juntas ontológicas corretas"* (1987, p. 130). Uma vez que tais juntas sejam encontradas, o FPIA poderá ser resolvido. Este seria um problema distinto do FP porque, como visto, Hayes queria estudar a inteligência como fenômeno amplo, sem preocupar-se com limitações psicológicas. Evidentemente, o recorte ontológico a que Hayes se referia precisaria ser levado em conta também em explicações dos mecanismos específicos por meio dos quais a mente humana realiza a inteligência, mas isto constituiria um problema distinto da FPIA. Em síntese, o FPIA seria o modo como o problema ontológico descrito se manifesta no âmbito das investigações acerca da inteligência ampla. Por sua vez, o FP seria o modo como este mesmo problema ontológico se manifesta no âmbito das investigações do

modo específico como a mente humana realiza a tal inteligência. Neste sentido, o FPIA não é um problema a ser resolvido pelas capacidades cognitivas humanas, mas sim dissolvido, ou seja, é preciso encontrar uma estrutura ontológica subjacente tal que faça com que o FPIA não possa vir à tona.

Contudo, esta tentativa de negar T2 sem negar T1 tem um problema: se o modo como Hayes caracteriza a distinção entre FP e FPIA tem ares de discussão meramente terminológica, é porque este é precisamente o caso. Ao expor o problema dessa forma, Hayes buscava combater o que considerava uma série de equívocos e impropriedades que alguns filósofos cometeram ao falar do FPIA.[25] Porém, muito deste esforço elucidativo terminou por reduzir-se a uma cruzada terminológica para evitar que o termo *"Frame Problem"* tivesse outro referente que não o FPIA. Como se verá, ao contrário do que pode parecer em um primeiro momento, a posição de Hayes não difere tanto assim da concepção filosófica do FP aqui apresentada. Por esta razão, pode-se afirmar que Hayes está sim comprometido com T2. Ele apenas reservaria o nome "*frame problem*" para o FPIA e daria ao FP um outro nome qualquer, como quando chegou a sugerir, de modo bastante sarcástico, o nome *"Fodor's problem"* (1987, p. 133).

De todo modo, foi para salvaguardar a possibilidade de posições metodologicamente distintas como esta que optou-se pela apresentação de T1 e T2 nestes moldes. Em síntese, se T2 for verdadeira, isto significa que a cognição humana, de algum modo, soluciona o FP. Resta explicar como. Por sua vez, a negação de T2 é a tese de que a cognição humana se constitui de um modo tal que o FP não tem a chance de vir à tona. [26]

Fodor é, possivelmente, o autor que opera os desdobramentos mais radicais a partir de sua defesa de T2 (1987). Assim como Dennett, ele descreve o FP em termos da dificuldade em explicar como é possível manter um sistema de crenças atualizado. Vem daí sua marcante analogia entre o FP e o problema de Hamlet: em que momento se pode dizer que um número "suficiente" de possibilidades foi considerado, permitindo encerrar o processo de revisão de crenças? Diante de um número potencialmente infinito de possibilidades, cada qual com seu conjunto de interdependências a

[25] Por vezes, este combate se dava de modo excessivamente ácido, como quando afirmou que *"Fodor não sabe a diferença entre o Frame Problem e um cacho de bananas"*. (1987, p. 132)

[26] Pode-se, claro, questionar: não seria tal distinção uma ilusão? Qual a diferença entre a afirmação de que a cognição humana resolve o FP, e a de que ela não tem problema algum a resolver? Toda e nenhuma. Tudo depende do modo como o FP é compreendido. Desta compreensão, vem aquilo que pode ou não ser aceito como uma solução adequada. Como já prenunciado, o FP foi apresentado de inúmeras formas por diversos autores. Para alguns deles, o FP é um problema inferencial cuja solução é essencialmente algorítmica. Ele nasce de pressuposições erradas acerca da natureza da mente humana. Se tais pressupostos forem abandonados, elimina-se a necessidade de um tratamento explícito do problema nos algoritmos e o FP deixa então de ser um problema para a cognição humana. Para outros autores, a natureza do FP vai mais fundo e impede o abandono destes pressupostos, a menos que se esteja disposto a abandonar a própria TCC. Como se verá, esta última será a posição para a qual se argumentará no decorrer dessa investigação: a única forma de negar T2 será negando a própria TCC. Contudo, neste estágio do argumento, em que ainda não se assumiu hipótese clara sobre a natureza do FP, é preciso salvaguardar a possibilidade de posições alternativas sobre.

serem testadas e consideradas, este é de fato um desafio monumental e, para Fodor, intransponível.

Ao colocar o FP deste modo, não é de se espantar que ele seja tão facilmente tomado como um problema constituído apenas por seu aspecto inferencial, despertando assim, além da ira terminológica de Hayes, o ceticismo de autores como McDermott de que haja aí de fato um problema insolúvel (mais sobre isto em seguida). Porém, não é justo acusar Fodor de reduzir o FP ao seu aspecto inferencial. Para ele, o desafio maior não é o de descobrir e implementar um procedimento algorítmico (tal como visto nas tentativas de solução para o FPIA) que possa fornecer um nível razoável de confiabilidade na delimitação das inferências a considerar. Em vez disso, o desafio exposto é o de modelar as capacidades cognitivas de modo a tornar possível uma explicação desta habilidade de encontrar critérios adequados em uma infinidade de diferentes contextos. Fodor se concentra não no aspecto inferencial, mas sim no que parece ser a contraparte filosófica do aspecto *ad hoc* dos axiomas de *frame*, tal como visto na apresentação do FPIA. Com efeito, para Fodor, tentativas de resolver somente o aspecto inferencial do FP sem atentar-se para o aspecto *ad hoc* dos critérios de relevância usados não geram soluções, mas reformulações do problema.

> O frame problem é tão ubíquo, tão polimorfo, e tão intimamente conectado com cada aspecto da tentativa de compreensão da inferência racional não demonstrativa, que é possível a um profissional falhar em perceber quando é, de fato, com o frame problem que ele está trabalhando. (FODOR, 1987, p. 142)

O que justifica uma afirmação tão contundente e geral acerca de um problema nascido num contexto tão circunscrito? A resposta está na natureza das dificuldades enfrentadas pelas tentativas de solução do FP. Conforme já visto, nenhuma delas obtém qualquer avanço para além do problema inicial, que é a demarcação dos elementos relevantes de um modo sensível a contexto. Todas, de algum modo, apenas deslocaram e reapresentaram o problema. O ponto de Fodor é que este aspecto das tentativas de solução nem sempre são evidentes para aqueles que as formularam. Tome-se como exemplo a abordagem *sleeping dog*. Parece possível programar um sistema de modo que implemente a LISC da seguinte maneira: nenhuma propriedade de nenhum objeto do micro-universo sofrerá qualquer modificação a menos que haja uma razão positiva para tal. De que modo isto se caracteriza como uma reformulação do problema? Como já exposto, o que conta como uma razão positiva para que uma dada inferência possa ser considerada relevante pode variar enormemente em função do contexto. Assim, uma razão positiva pode não se seguir imediatamente do conjunto de axiomas que descreve explicitamente o estado do sistema cognitivo. Pode ser necessário realizar um grande número de derivações até que se encontre uma inferência que possa ser tomada como uma razão positiva. Tal número de derivações necessárias pode, facilmente, tornar-se intratável, visto que pode haver, inclusive, razões que serão consideradas positivas, mas que estão ocultas na interdependência

entre inferências no interior do universo considerado. Por fim, pode ser também o caso de que não haja razão positiva alguma, isto é, que a conclusão mais adequada seja a de supor que tudo se mantém tal como o modo *default* da LISC supõe em seu ponto de partida. Porém, quantos testes devem ser feitos até que esta se mostre a conclusão mais razoável? O problema de Hamlet, portanto, permanece: quando parar de pensar?

O acúmulo de casos como este faz com que Fodor sinta-se justificado em adotar uma indução pessimista e afirmar que qualquer tentativa de solucionar o FP tendo apenas seu aspecto inferencial como foco resultará numa reformulação do problema. Este é um efeito que se obtém em toda tentativa de calcular a relevância de uma crença ou fato: não parece possível encontrar um critério de parada ou de suficiência que não seja arbitrário sem antes checar todas as possibilidades do universo em questão. Para Fodor, o foco no aspecto inferencial desvia a atenção do caráter *ad hoc* de toda tentativa de estabelecer princípios utilizados para circunscrever elementos relevantes em processos cognitivos. Ignorar este aspecto resulta na sensação enganosa de que o FP pode ser solucionado, tendo como efeito apenas uma reformulação do problema. Se o caráter *ad hoc* for reconhecido como uma característica de qualquer princípio geral que se tente adotar, então será simples reconhecer o FP como um problema computacionalmente insolúvel.

Contudo, a motivação de Fodor não se restringe a esta indução pessimista. Sua argumentação tem como pano de fundo a sua tese da modularidade da mente (1983). Segundo esta tese, a mente possui módulos dedicados a tarefas específicas, tais como os sistemas perceptuais e/ou subsistemas específicos deste (reconhecimento facial, por exemplo). Fodor toma o *encapsulamento informacional* como uma característica fundamental destes módulos. O significado preciso deste termo pode ser objeto de disputa, mas para os fins desta investigação, pode-se aceitar que um processo é dito informacionalmente encapsulado quando ele só tem acesso a um tipo próprio de informação, e não informações de outros módulos. Ilusões de ótica constituem os casos mais clássicos e simples do que Fodor tem em mente. Tome-se como exemplo a ilusão de Müller-Lyer:

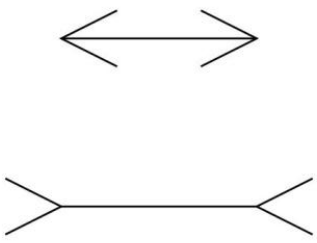

Figura 2 – Ilusão de Müller-Lyer

Embora as duas linhas horizontais possuam largura idêntica, há uma forte impressão de que uma é mais larga que a outra. Essa impressão persiste mesmo diante

da informação de que trata-se de uma ilusão, e mesmo que se faça uso de uma régua ou aparato similar para checar a largura das linhas. Na tese de Fodor, isto se explica porque o processamento da imagem ocorre num módulo informacionalmente encapsulado que não faz uso de quaisquer informações obtidas por outros meios, isto é, por outros módulos.

Por sua vez, processos que não possuem encapsulamento informacional são chamados *isotrópicos*. Nestes processos, a ausência de encapsulamento *a priori* faz com que todo e qualquer fato ou crença disponível no domínio cognitivo seja potencialmente relevante quando avaliando qualquer outro fato ou crença. *Prima facie*, não parece haver relação alguma entre o resultado do torneio mundial de xadrez e a possibilidade do cachorro do vizinho se assustar no próximo sábado pela manhã. Porém, dado um conjunto de crenças de fundo que inclua elementos como: há um segundo vizinho cujo irmão está participando da final do torneio mundial de xadrez, este segundo vizinho costuma usar fogos de artifício em ocasiões comemorativas, e caso seu irmão vença, provavelmente soltará alguns fogos. Além disso, a partida está marcada para o próximo sábado pela manhã e, ao ouvir fogos de artifício, o cachorro do primeiro vizinho tende a assustar-se. Diante disso, é plausível afirmar que, no contexto exposto, crenças sobre o resultado do torneio mundial de xadrez são relevantes para crenças sobre a possibilidade de o cachorro de um vizinho assustar-se.

Assim, um processo cognitivo idealizado é, sem dúvida, isotrópico, e este parece ser o caso dos processos que pautavam as pesquisas em busca do algoritmo da inteligência ideal, mas é este o caso da cognição humana? De modo mais específico: será o caso que a determinação de elementos relevantes em um dado domínio cognitivo é um exemplo de processo isotrópico? Para Fodor, não há dúvida que sim. Um dos exemplos que utiliza para ilustrar esse ponto é o do raciocínio por analogia: *"...um processo que depende precisamente da transferência de informação entre domínios cognitivos previamente assumidos como mutuamente irrelevantes"* (1983, p. 107). Deste modo, Fodor acredita haver uma dicotomia entre ser informacionalmente encapsulado e apresentar o FP: *"ser inteligente, não modular e fazer surgir o frame problem vêm sempre em conjunto"* (1987, p. 141). Em conclusão, somente sistemas informacionalmente encapsulados são computacionalmente tratáveis, podendo evitar problemas de relevância como o FP.

Não é necessário avaliar em maior profundidade a pertinência dos argumentos de Fodor acerca da impossibilidade de uma teoria para processos cognitivos não informacionalmente encapsulados. Importa apenas notar que, para além da indução possibilitada pelo fracasso em solucionar ou dissolver o FP, Fodor apresenta também razões conceituais para pensar que não é possível achar nenhuma base firme *a priori* para solucioná-lo (2001). Assim, quaisquer tentativas de descrever ou explicar processos isotrópicos em termos computacionais originarão, elas mesmas, uma nova versão do FP. Para Fodor, é precisamente isto o que acontece em tentativas de solução por meio das abordagens *cheap tests* e *sleeping dog*.

Tem-se assim um cenário em que T2 não só é tomada como verdadeira, como também possui consequências graves para qualquer empreitada científica que busque compreender processos cognitivos sem encapsulamento informacional, tal como, ao menos para Fodor, é o caso da razão humana. De fato, Fodor rejeita a possibilidade de que tais processos possam ser objeto de estudo de uma ciência cognitiva, e esta seria uma limitação da própria TCC. Como para Fodor não há nenhuma outra teoria plausível, ele não hesita em fazer desta questão conceitual uma fonte de forte constrangimento das possibilidades de pesquisa empírica. Nesse sentido, o FP seria um problema não apenas um problema para o engenheiro da GOFAI, ou mesmo para o cientista cognitivo, mas para qualquer um que participe, de algum modo, do paradigma cognitivista. Essa crença fez com que Fodor chegasse a afirmar o FP é *"... importante demais para ser deixado nas mãos dos hackers."* (1987, p. 148), enfurecendo Hayes, mas enfatizando o caráter filosófico da empreitada.

Evidentemente, as bases para uma tal afirmação podem ser questionadas de diversas formas. É possível, por exemplo, rejeitar a tese de que a cognição humana envolva processos isotrópicos e, ainda assim, entender que T2 é verdadeira. Antes de considerar posições intermediárias, contudo, convém compreender as razões que podem levar alguém a defender uma posição diametralmente oposta: a rejeição de T2 pela negação de que o FP seja um problema real.

1.4.2 Heurística e formalismos não monotônicos: quão real é o problema?

Durante a apresentação do *Unless*, foi mencionada brevemente uma característica da lógica clássica que poderia estar relacionada ao FPIA: a monotonicidade. Em parte, a aposta de Sandewall era de que o uso de um formalismo não monotônico poderia evitar o FPIA. Como visto, o problema é que formalismos monotônicos permitem representar fatos, mas a LISC não é um fato, e sim uma cláusula *ceteris paribus*, uma espécie de modo *default* de raciocínio. Com efeito, a substituição do cálculo de situação por um outro tipo de formalismo não restrito às características basilares da lógica clássica abre novos caminhos para realizar cálculos que respeitem uma cláusula *ceteris paribus* específica para cada cenário. Intuitivamente, isto parece tornar mais concreta a possibilidade de implementação da LISC. Dennett, por exemplo, acredita que formalismos não monotônicos parecem harmonizar melhor com o modo tal como a inteligência humana funciona (1987, p. 56).

Em virtude desta possibilidade, houve quem negasse que o FPIA fosse um problema real para a GOFAI. McDermott (1987) argumenta que o FPIA é um problema restrito ao uso do cálculo de situação e sua inverossímil tentativa de modelar a dinamicidade do universo real nos termos de uma sequência de micro-ambientes gerada por um único agente, que é o próprio sistema. O FPIA não é um problema de caráter ontológico, no sentido exposto por Hayes, mas sim fruto do uso de um formalismo

inadequado. Não é um desafio posto pelo objeto de pesquisa (a inteligência) mas sim uma dificuldade oriunda de uma ferramenta inadequada.

Como visto anteriormente, o FPIA tal como manifesto no contexto do cálculo de situação apresenta dois desafios: *descrever* um número potencialmente infinito de axiomas de *frame* (o chamado aspecto *ad hoc*), e *calcular* os não efeitos a partir deles (o chamado aspecto inferencial). A tese de McDermott é que o uso de formalismos não monotônicos, tais como a *lógica não monotônica*, de sua autoria (1980), em conjunção com alguma variante da estratégia *sleeping dog*, constituem uma ferramenta adequada para evitar o aspecto *ad hoc* das outras soluções e reduzir o FPIA ao seu aspecto inferencial. Embora não seja mais necessário especificar os axiomas de *frame*, continua sendo necessário calcular os não efeitos de ações, e a magnitude do cálculo continua proporcional ao número de possíveis efeitos. McDermott descreve este aspecto como o sendo o problema de: *"..inferir de modo eficiente que um fato ainda é verdadeiro em uma situação"* (1987, p. 115). Trata-se do problema anteriormente apontado no âmbito da estratégia *sleeping dog*: se para cada evento envolvendo um fato básico for necessário percorrer a lista completa de todos os possíveis fatos não básicos que dele podem ser derivados, então os cálculos necessários, mesmo para eventos simples, tenderão ao gigantismo. Para McDermott, este é um problema difícil, mas não há motivo para pensar que ele seja insolúvel ou que seja responsável pela queda da GOFAI.

Exposta a posição de McDermott sobre o FPIA, resta ver sua compreensão acerca do FP, isto é, da contraparte filosófica tal como detectada por Denett e expressa por Haugeland. McDermott recusa tanto que o FPIA seja um problema para sistemas computacionais, quanto que o FP seja um problema epistemológico para seres humanos: *"nunca vi um exemplo convincente, sendo que todas as propostas eram ou fáceis demais para máquinas ou difíceis demais para qualquer um, inclusive humanos"* (1987, p. 117). Sua motivação envolve, além dos elementos já citados, a possibilidade de usar modelos imperfeitos e *heurística*. Tal possibilidade já foi vislumbrada durante a discussão do STRIPS. McDermott, contudo, não é defensor do STRIPS, que usa um formalismo monotônico. Ele defende que o uso de formalismos não monotônicos é o que permite que a abordagem *sleeping dog* tenha sucesso.[27]

Em um jogo de xadrez, na impossibilidade de checar todas as possíveis consequências a fim de optar pelo melhor movimento, pode-se guiar a decisão por uma regra que tipicamente gera bons resultados, tal como ampliar o domínio no centro do tabuleiro. Responder à pergunta "qual a jogada que mais amplia o domínio do centro?" é muito mais simples e envolve um número substancialmente menor de possibilidades que a pergunta "qual a melhor jogada possível?". Evidentemente, a melhor jogada possível pode não envolver nenhuma tentativa de ampliar o domínio do centro

[27] Isto não significa que ele defenda a abordagem de Sandewall em particular. McDermott prefere manter o foco em seu próprio formalismo (1980), ou discutir outros, como o *circunscription* de MCCARTHY (1980).

do tabuleiro. Pode ser o momento perfeito para sacrificar uma peça por outra, mas a adoção da heurística fará com que esta opção sequer seja considerada. Assim, fazer uso de heurística significa fazer uso de inferências não dedutivas que podem falhar por deixar de fazer inferências que seriam "óbvias" de um ponto de vista estritamente formal. McDermott não vê problema na possibilidade de falha, visto que seres humanos também falham, ou seja, seres humanos deixam de fazer inferências "óbvias" o tempo todo. Ele mesmo descreve um cenário em que a descarga da máquina de lavar roupas apontava para o mesmo tanque em que ele havia deixado algumas roupas de molho (1987, p. 113). Quando a descarga começou, a inundação foi inevitável. Como explicar, ao mesmo tempo, a falha em realizar uma inferência tão simples (realizar a descarga de água da máquina de lavar roupas em um tanque cheio o fará transbordar) e a capacidade de explicar imediatamente o que aconteceu, uma vez que se presencie o fato? Simples: o limite das opções consideradas pode ser estabelecido por modelos incompletos, e o grau de completude suficiente pode ser estabelecido por meio de heurísticas. Com efeito, o aspecto *ad hoc* do FPIA é um falso problema, pois não se faz necessária a infalibilidade, mas apenas a obtenção de um grau de confiabilidade paralelo ao que seres humanos tipicamente apresentam.

Isso coloca McDermott também contra Hayes, pois ele defende que, caso exista um problema tal qual o FP, ele não é um problema da GOFAI mas sim dos filósofos que o adotaram, e somente na versão em que o adotaram. Desta perspectiva, críticas como as de Haugeland (1987) à abordagem *sleeping dog*, por exemplo, erram o alvo porque exigem infalibilidade e não apenas confiabilidade. Ainda que, no discurso, estes argumentem que não estão a exigir infalibilidade, para McDermott, este é um pressuposto implícitos nos argumentos utilizados. Assim, o caminho para a solução do problema inferencial manifesto está dado: basta desenvolver modelos e heurísticas que permitam obter um grau de acerto e erro que coincida com o do ser humano.

Em síntese, enquanto McDermott acredita que o uso de formalismos não monotônicos, em conjunção com uma estratégia *sleeping dog*, pode evitar que o FP venha à tona, Fodor acredita que o resultado desta estratégia é apenas uma formulação diferente do mesmo problema. No coração da discordância estão concepções distintas da natureza do FP. Esta diferença pode ser elucidada a partir de uma análise mais minuciosa do papel e dos limites que cada um dos autores concede ao uso de heurística. A tese a que McDermott subscreve pode ser sintetizada da seguinte forma:

Tese 3 (T3): Reconhecer que buscamos apenas confiabilidade, e não infalibidade, faz do FP um falso problema para a cognição humana.

Qual seria a posição de Fodor diante de T3? Ao contrário do que McDermott faz parecer, Fodor não exige mais do que a obtenção de resultados geralmente confiáveis quando o que está em xeque é o aspecto inferencial do FP. Fodor não está a exigir infalibilidade de um sistema cognitivo. Seu ponto é outro: uma solução para o FP não pode envolver elementos *ad hoc* ou arbitrários. Para Fodor, isto faz com que

se demande mais do que a conjunção de heurística, formalismos não monotônicos e modelos incompletos. É preciso justificar o modelo em particular e o conjunto de heurísticas que se quer adotar. Porém, esta justificativa invariavelmente oscilará entre o *ad hoc* e uma reformulação do FP. É evidente, portanto, que Fodor rejeita T3. Ele considera este apelo às limitações humanas uma justificativa infundada para o uso de heurística em processos de determinação de relevância. O uso de heurística está entre *"... as formas mais características que os cientistas cognitivos tem de não reconhecer o quão sérios são os problemas que a isotropia da relevância traz à tona para teorias da mente cognitiva."* (2010, p. 116).

O problema se manifesta do seguinte modo: antes de fazer uso de uma regra de heurística, o sistema cognitivo precisa *decidir* qual é a regra heurística adequada para a situação em que o sistema se encontra. Voltando ao exemplo do jogo de xadrez: em quais situações do jogo é adequado buscar ampliar o domínio do centro do tabuleiro? Isto é, em que momentos deve-se fazer uso de tal regra heurística? A resposta mais comum entre os defensores da abordagem *sleeping dog*, por exemplo, é dizer que a regra deve ser utilizada nas situações em que ela se mostrou previamente confiável. Recorre-se assim à familiaridade e experiência prévias. Em geral, a busca do domínio do centro do tabuleiro é algo pelo qual se nortear logo após o fim das aberturas iniciais dos jogadores. O alvo de Fodor, contudo, é precisamente o mecanismo utilizado para identificar as situações adequadas: *"... frequentemente, o problema da relevância vem à tona ao tomar por dado algum princípio para individuar situações; um princípio tal que ele mesmo pressuponha uma noção de relevância completamente inexplicada."* (2010, p. 118). Este processo decisório, como Fodor nota, é isotrópico, pois depende da detecção prévia de uma situação como sendo uma situação de um determinado tipo, e isto depende da relevância dos elementos presentes no contexto, mesmo que o objetivo seja reconhecer uma dada situação como uma situação na qual já se esteve antes. Vale ressaltar como o ponto de Fodor pode ser ilustrado pelos resultados obtidos a partir das abordagens *cheap test* e *sleeping dog*: a demanda por categorizações de ordem superior (como estereótipos), e os problemas que daí advém.

Contudo, pode-se tentar resistir à argumentação de Fodor do seguinte modo: nem mesmo a escolha da heurística adequada precisa ser perfeita, mas apenas suficientemente confiável. Tal afrouxamento tem consequências sobre o que aparece como computacionalmente tratável ou não. Para Fodor, somente processos informacionalmente encapsulados (isto é, não isotrópicos) são tratáveis, visto que, nestes casos, o subconjunto dos elementos cognitivos a se considerar é um dado, e não um elemento a ser computado. Se a delimitação das informações relevantes for, ela mesma, parte do processo cognitivo, o resultado é o retorno do problema: a delimitação é, ela mesma, computacionalmente intratável, afinal, o processo precisará checar todas as possibilidades antes de decidir quais delas são ou não relevantes. Quem articula uma resposta mais direta a isto é SAMUELS (2010): não é necessário que a heurística seja capaz de circunscrever os elementos efetivamente relevantes tal como o faria

uma limitação pré-determinada no domínio cognitivo. Basta que a heurística funcione como uma forma de restringir o escopo daquilo que será considerado no processo inferencial de um modo que apresente suficiente coincidência com os elementos tipicamente relevantes nos cenários considerados. O que Samuels sugere é que o encapsulamento informacional é condição suficiente, mas não necessária, para que um processo seja computacionalmente tratável. De outro modo: todo processo informacionalmente encapsulado é computacionalmente tratável, mas nem todo processo computacionalmente tratável é informacionalmente encapsulado.

Para Samuels, técnicas que envolvem o uso de heurística permitem que um processo circunscreva, ele mesmo, o volume de informações consideradas, fazendo dele um processo computacionalmente tratável, ainda que não seja informacionalmente encapsulado. Se a circunscrição no volume de informações apresentar suficiente coincidência com o conjunto de elementos relevantes para a tarefa, então o processo será computacionalmente tratável e confiável. A operação será, sem dúvida, falível, mas como lembra McDermott, também o são os processos cognitivos do ser humano. A circunscrição não precisa demonstrar acurácia invariável.

Em sua argumentação, Samuels apela para uma analogia com o tipo de estratégia utilizada em sistemas de busca contemporâneos como o Google. Quando se realiza uma busca, os resultados obtidos parecem demonstrar uma capacidade razoável de discernir informações relevantes em meio a uma quantidade massiva de dados, sem qualquer encapsulamento *a priori*. Samuels não qualifica o que exatamente ele considera elucidativo no uso desta analogia, mas a resposta mais provável é que sistemas de busca fazem uso de heurística para determinar as ocorrências mais relevantes dos termos buscados.

Supondo que alguém faça uma busca pelo termo “barcelona”, como o sistema decide quais são as ocorrências mais relevantes? Boa parte das heurísticas aplicadas ao realizar a circunscrição das informações levam em conta o perfil do indivíduo: se ele gosta de futebol, as páginas ligadas ao time espanhol podem ter mais relevância. Se ele gosta de viajar, sites com informações turísticas podem ser priorizados. Do mesmo modo, um perfil acadêmico pode fazer com que informações acerca da Universidade de Barcelona façam parte do conjunto de elementos relevantes. O processo é falível (mesmo quem gosta de futebol pode querer viajar ou estudar em Barcelona) mas suficientemente confiável. Contudo, nem sempre o perfil está disponível. Como se dá a delimitação nestes casos? Mais heurística: páginas com mais menções do termo são mais relevantes que páginas com menor número de menções; páginas em que o termo apareça no título são mais relevantes que páginas em que o termo só apareça no meio de um texto; páginas de sites de reconhecida qualidade são mais relevantes do que sites obscuros, e assim por diante. *Prima facie*, a proposta de Samuels soa plausível: não poderia algo análogo se dar com os processos cognitivos do ser humano?

O problema com esta estratégia é que ela parece errar o alvo. A tentativa de negar a gravidade do FP termina por enfatizar apenas seu aspecto inferencial. O vago

apelo de Samuels às técnicas utilizadas por sistemas de busca, por exemplo, ignoram o fato de que tais algoritmos têm sim condições de delimitar previamente o escopo em que as buscas se realizam. Isso é possível porque eles atuam não sobre o universo real, cujas "juntas ontológicas" (como diria Hayes) são desconhecidas, mas sobre bases de dados previamente organizadas para fins delimitados. A heurística é eficiente porque a informação foi previamente organizada para servi-la. Assim, a circunscrição daquilo que é relevante para uma determinada busca é calculável a partir de parâmetros fixos definidos *a priori* pelos próprios programadores ao determinar quais as propriedades das informações que serão armazenadas. A atribuição de parâmetros fixos análogos a um sistema cognitivo que vá lidar com o universo real constituiria uma tentativa de solução caracteristicamente *ad hoc*. Portanto, um tal apelo à heurística não resolve, mas antes *pressupõe* uma solução prévia para a circunscrição do domínio sobre o qual o sistema atuará.

A questão organizacional acima esboçada é de fundamental importância para a caracterização do FP, e será melhor desenvolvida em breve. Antes, é preciso elaborar um pouco mais o contraste entre a posição de Fodor e Samuels. Isto permitirá compreender por que certas críticas de Samuels podem aplicar-se à Fodor, mas não à versão do FP aqui apresentada. Este embate entre as posições de Fodor e Samuels em torno de T3 adquire nova roupagem quando Fodor introduz a noção de *propriedades globais* (2001). Uma propriedade global é uma propriedade que só pode ser atribuída a uma representação (crença, fato etc.) em função de sua relação com todas as demais representações no sistema.[28] A relevância, para Fodor, é um exemplo. Uma crença não é intrinsecamente relevante: ela pode vir a ser relevante, a depender das circunstâncias e das relações que mantém com a totalidade do conjunto de crenças do agente.

A noção de propriedade global é utilizada para destacar o contraste destas com a *localidade* dos processos computacionais. Fodor argumenta que somente propriedades sintáticas são acessíveis aos processos computacionais.[29] Isso significa que o papel causal de uma representação em um processo computacional é dado exclusivamente por sua forma, ou seja, é fruto de sua estrutura constitutiva. Expressar as propriedades locais de uma representação é descrever suas propriedades sintáticas. Fodor diz: *"se uma certa representação mental está ou não no domínio de um certo processo mental depende exclusivamente das relações entre esta representação e suas partes"* (2010, p. 108).

Quando a realização de um processo cognitivo é dependente de propriedades

[28] É importante enfatizar que, na concepção de Fodor, a propriedade global de uma representação é fruto da sua relação com todas, e não somente algumas representações presentes no sistema. De todo modo, uma leitura segundo a qual uma propriedade global guarda relação com algumas, mas não necessariamente todas, as representações presentes no sistema não passaria de uma reformulação do FP, afinal, seria preciso agora mapear quais, dentre todas elas, detém ou não tal relação. Mesmo assim, há quem tente seguir este caminho, como será discutido no capítulo 3.

[29] Trata-se de uma suposição que vem deste Turing, mas que vem sendo questionada, embora de modo ainda incipiente, por autores como RESCORLA (2017).

globais, ele é denominado *processo global*. Dizer de um processo que ele é global é mais do que dizer que ele é isotrópico. Embora ser isotrópico e ser global sejam características intimamente relacionadas, elas não são idênticas. Se um sistema de crenças é isotrópico, qualquer dos fatos contidos em seu domínio pode vir a ser relevante para um processo cognitivo. Contudo, como esclarece Samuels: *"não é meramente o caso que crenças relevantes podem vir de qualquer lugar, mas sim que elas rotineiramente vem de (quase) toda parte."* (2010, p. 285). Se um processo é global, isto significa que mesmo as tarefas mais simples, em geral, demandam considerações envolvendo grande número de crenças a considerar. Fodor defende a tese de que a maior parte dos processos mentais (virtualmente todos os processos de revisão de crenças, por exemplo) são globais. Assim, a relevância é uma propriedade global indisponível para processos computacionais. Trata-se não da exceção, mas da regra. Porém, se os processos cognitivos forem computacionais, estes serão locais. A existência de propriedades globais que participam dos processos cognitivos pode ser tomada como uma reformulação do argumento pela impossibilidade de uma ciência cognitiva que abarque tais processos. O resultado é uma nova roupagem para o argumento que Fodor usara previamente para o mesmo efeito, e pode ser sintetizado do seguinte modo:

(1) a cognição humana depende do uso de propriedades globais mesmo para tarefas que realiza rotineiramente, tais como a circunscrição de elementos relevantes;
(2) propriedades globais não são computáveis; logo
(3) a cognição humana não pode ser explicada computacionalmente.

Assim, ele conclui: *"...parece claro que, seja qual for a solução que o frame problem venha a ter, ela não será computacionalmente local"* (2010, p. 121). Ora, uma solução computacionalmente local é precisamente o que Samuels, McDermott e outros acreditam ser possível por meio de heurística. Com esta linha de argumentação, Fodor tenta mostrar que tal projeto está fadado ao fracasso. Mesmo a seleção da heurística adequada é um processo global que repousa sobre a necessidade de considerar todas as possibilidades. Como as experiências com abordagens *cheap test* e *sleeping dog* demonstraram, tentativas de evitar esta necessidade de acessar propriedades globais terminam por mostrar-se meras reformulações do problema.

A esta altura, não é difícil notar que Samuels não tem motivo para se deixar pressionar por esta reformulação. Ele pode resistir à conclusão como vem fazendo até aqui: concordando com Fodor que propriedades globais não são computacionalmente acessíveis, mas negando (1), isto é, negando que tais propriedades sejam efetivamente acessadas pelos processos cognitivos humanos. O que há são processos que conseguem atingir suficiente coincidência com o que seria o caso se a cognição humana tivesse de fato acesso a propriedades globais. Assim, a estratégia de Samuels é, novamente, enfraquecer os requisitos necessários para que um processo cognitivo que envolva propriedades globais possa ser considerado computável.

Em sua interpretação de Fodor, Samuels nota que ele caracteriza a posse de uma propriedade global (como a de ser relevante) como equivalente a uma certa caracterização da sensibilidade ao contexto (2010). Explicar como é possível que a cognição humana seja sensível ao contexto demanda explicar como é possível que ela acesse propriedades globais, e isto está fora do alcance da TCC. Contudo, Samuels argumenta, esta caracterização não descreve necessariamente o modo como o ser humano de fato pensa. Trata-se, antes, de uma forma de determinar normativamente os requisitos necessários para uma razão idealizada, que estaria presente em um agente idealizado, e não da descrição de um feito constante da cognição humana. Samuels usa o exemplo da consistência como o tipo de padrão normativo pelo qual um conjunto de crenças pode ser analisado. Uma crença tem a propriedade global de ser consistente quando é consistente com a totalidade do conjunto de crenças do agente. Contudo, parece fora de questão que o ser humano seja um poço de inconsistências, ainda que a busca pela consistência guie os processos que envolvem subconjuntos de crenças. A relação entre a consistência e os processos cognitivos da mente humana é, nesse sentido, normativa. O ser humano é um agente que não consegue alcançar plenamente esse ideal, mas apenas de modo parcial e falível. Assim, Samuels concorda com Fodor que relevância e consistência são propriedades globais, mas discorda que disto se siga a inexplicabilidade dos mecanismos por meio dos quais elas se dão.[30] É preciso então fazer aterrissar as demandas de Fodor. Em vez de exigir dos processos computacionais algo que não podem entregar, pode-se afrouxar o requisitos daquilo que se considera um aparato necessário ao tratamento adequado dos processos mentais, de modo a torná-los mais plausíveis para criaturas finitas como o ser humano.

O que se pode concluir diante desta discussão acerca de T3? Pode o FP ser evitado por meio da adoção de diferentes técnicas e formalismos como querem Fikes (STRIPS), Sandewall e McDermott? Ou são tais tentativas meras reformulações do problema, meras formas de ocultar o caráter *ad hoc* das soluções, como sugere Fodor? Antes de responder esta pergunta, é preciso colocar ordem na casa. A esta altura, espera-se ter deixado claro que boa parte da discussão apresentada, desde as tentativas de solução para o FPIA até as formulações de Fodor, McDermott e Samuels, tratam das mesmas questões, mas sob diversas roupagens.

Um aspecto comum a quase toda argumentação acerca do FP e do FPIA é que ela enfatiza seu *aspecto inferencial*. Abordagens *cheap test* buscam reduzir o número de testes para realizar inferências relevantes. Abordagens *sleeping dog* (grupo que inclui os projetos de McDermott) buscam eliminar a necessidade da maior parte dos testes, mantendo assim o foco nas inferências relevantes. Samuels acredita que o segredo está em encontrar uma estratégia adequada para realizar inferências. Em todos estes casos, há uma ênfase na busca por uma estratégia algorítmica que permita superar o desafio posto pelo FP. Tal ênfase pode ser encontrada mesmo em Fodor.

[30] É possível também argumentar que a relevância não é uma propriedade global. Esta questão em particular será retomada e melhor desenvolvida no terceiro capítulo.

Apesar de este enaltecer o aspecto *ad hoc* das soluções, ele não tenta elucidar a natureza daquilo que faz com que estas tentativas de solução tenham esta característica. Em vez disso, ele faz uma indução pessimista e conclui que não há solução possível, o que é fatal para a TCC. Fodor pode estar certo em dizer que não há solução possível, mas tal afirmação demanda melhor qualificação. É preciso, portanto, um esforço para compreender em função de que as tentativas de solução oscilam entre reformulações do problema ou estratégias *ad hoc*.

A expectativa é de que a elucidação deste ponto permita defender que, contra Samuels e McDermott, Fodor tem razão ao dizer que o FP não pode ser resolvido com técnicas envolvendo heurística, ainda que não tenha razão em defender que a sensibilidade ao contexto se dá pelo acesso a propriedades globais. De outro modo: enunciar uma solução computacionalmente tratável para a sensibilidade ao contexto não é o mesmo que enunciar uma solução computacionalmente tratável para o acesso a propriedades globais. Ao contrário do que talvez possa parecer, isto não demanda nenhum novo *insight*, mas sim o resgate de um aspecto do problema que não foi suficientemente enfatizado quando de sua migração para o meio filosófico. Uma solução adequada não lida apenas com o modo pelo qual a informação é acessada ou navegada, mas também com o modo como ela é modelada. Ao desenvolver este aspecto, será possível perceber melhor por que o FP é tão resiliente.

1.5 O aspecto organizacional

A discussão até aqui apresentada permite notar que o FP é um problema de relevância que, embora descrito sempre de modo ligado à revisão de crenças, vai muito além disso. Ele aparece em todo tipo de tarefa cognitiva, tais como planejamento, tomada de decisão e interpretação de conteúdo perceptual. Além disso, aparece também em quaisquer empreitadas racionais idealizadas que possuam o modelo da inferência não demonstrativas, tal como a própria empreitada científica. Atento a isto, GLYMOUR (1987) tenta mostrar que tanto o FP quanto o FPIA não constituem um problema novo, mas sim uma nova roupagem para um problema antigo, ou melhor, para uma família de antigos problemas relacionados à relevância. A novidade estaria apenas nas condições em que estes problemas são apreciados. Ele diz: *"todas as instâncias do frame problem tem a forma: (. . .) dada uma quantidade enorme de coisas e, dada uma tarefa para ser feita usando algumas destas coisas, quais são as coisas relevantes para a tarefa?"* (1987, p. 65). O que Glymour sugere, talvez sem perceber, é que se faça uma inversão metodológica no modo como o FP é abordado. Discutir o problema a partir de empreitadas específicas, seja a questão original da abordagem dedutivista na GOFAI, seja a questão epistemológica colocada por Dennett, faz com que o debate seja poluído por detalhes específicos das respectivas empreitadas, além de contribuir para a confusão conceitual que caracteriza boa parte da literatura sobre

o tema. Nesse sentido, pode ser mais produtivo assumir um “marco zero” e buscar, a partir dele, as condições mínimas necessárias para que o FP se mostre. Esse marco zero, sugere-se aqui, pode ser constituído a partir da descrição de Glymour e denominado *Problema Geral da Relevância* (PGR). O FP será então caracterizado como uma versão ou espécie do PGR.

Glymour especifica também dois requisitos que caracterizam o modo como o PGR foi tomado por Dennett ao fazer do FP uma preocupação filosófica: primeiro, o *requisito computacional*: uma solução adequada ao PGR precisa ser computacionalmente tratável. Segundo, o *requisito antropocêntrico*: uma solução adequada precisa lidar com o PGR de modo análogo ao modo como os seres humanos o fazem.[31] Tais requisitos podem ser compreendidos como critérios de especiação sobre um gênero comum, que é o próprio PGR. Por ora, ficará em aberto se tais requisitos de fato descrevem de modo suficientemente preciso as condições na qual o FP se manifesta.

Tem-se então uma ferramenta que permitirá expressar melhor o que há de comum, e o que há de específico aos problemas ligados à relevância. Uma descrição do FPIA pode, por exemplo, demandar requisitos adicionais, tais como um *requisito lógico*. Isto significa que a solução ao PGR em sua versão FPIA precisa ser não apenas computável, mas também implementável segundo uma lógica dedutivista de primeira ordem. Porém, como se verá, não é tão óbvio que mesmo este requisito esgote ou caracterize adequadamente o cenário em que o FPIA surgiu. Podem estar em jogo também pressupostos do cognitivismo, concepções de inteligência e até mesmo subscrição à teses como a de que todo tipo a lógica clássica é capaz de modelar qualquer coisa, tese defendida por HAYES (1977).

O importante, por ora, é compreender um conjunto de requisitos como aquilo que atribui concretude ao PGR, permitindo tomá-lo como um problema em particular, cujas condições para uma solução satisfatória podem ser explicitadas de modo claro. O PGR é, por si só, um problema geral e vago demais para ser articulado de maneira útil. Com efeito, a discussão sobre o conjunto de requisitos a que o PGR é submetido é tão importante quanto a busca por uma solução, uma vez que a falha em reconhecer requisitos articulados, ou mesmo uma falha em reconhecer efeitos da conjunção de diferentes requisitos pode fazer com que uma versão do PGR como o FP seja (ou não) insolúvel. Note-se, porém, que ao estruturar a investigação desta forma, não se busca afirmar aqui que o PGR é um único e grande problema, como se a solução definitiva para qualquer de suas manifestações, junto à qualquer conjunto de requisitos, demandasse uma solução geral para o próprio PGR. Trata-se antes de uma ferramenta metodológica para organizar concepções de manifestações concretas do PGR.

[31] Note-se que o requisito antropocêntrico está presente em toda caracterização do FP, ainda que não esteja presente em toda caracterização do FPIA. Como visto, havia linhas de pesquisa que se pautavam pelo estudo da inteligência sem preocupação com questões psicológicas. No campo filosófico, porém, este nunca foi o caso.

Um rápido exemplo do uso deste ferramental pode ser elucidativo. Como visto anteriormente, Samuels recorre ao caso dos sistemas de busca contemporâneos para exemplificar o modo como é possível lidar com uma quantidade maciça e crescente de informações sem qualquer tipo de encapsulamento informacional determinado *a priori*. Contudo, esta analogia falha, e as razões da falha podem ser expressas em termos de diferentes conjuntos de requisitos a que cada empreitada está submetida. Embora nos dois casos haja um requisito computacional, a quantidade de recursos disponíveis para um ser humano é substancialmente inferior à quantidade de recursos considerada razoável para uma empreitada tal como a de prover um sistema de busca eficiente. O sistema de busca não está, portanto, submetido a um requisito antropológico do tipo demandado para uma solução satisfatória do FP.

Embora tal assimetria não pareça insuperável, é possível diagnosticar uma diferença mais profunda que pode ser denominada *requisito ontológico*:[32] enquanto um ser humano habita um universo real, um sistema de busca "habita" apenas um conjunto de dados previamente desenhado para facilitar ao máximo as operações de busca. As decisões sobre que tipo de objeto será representado, bem como de quais serão suas propriedades e relações, foi uma decisão de *design*, e as limitações dos serviços oferecidos são levadas em conta na hora de fazer uso do sistema (até o presente momento, salvo de modo limitado, não há sistema que permita realizar buscas como "fotos em que meu cantor favorito esteja usando camisa amarela"). Além disso, os limites da ontologia do sistema de busca podem ser também modulados por diversos outros fatores, inclusive interesses humanos de ordem comercial ou política, que podem vir a constituir requisitos adicionais. Fatores como estes ditam ao sistema aquilo que é ou não relevante representar e o modo como esta representação é estruturada. Assim, o sistema de busca precisa lidar com um "mundo" (uma base de dados) reduzido e estruturado de modo a viabilizar certos serviços disponibilizados aos seus usuários, e não com o universo *real*. Como a GOFAI bem sabe desde seus primórdios, criar um micro universo e realizar tarefas específicas nele é muito mais fácil que desvendar a ontologia do universo real. Apesar do volume de informações envolvido, trata-se de um micro universo, tal como no caso de DR31FUS.

Resta agora estabelecer qual a versão do PGR que será de interesse no restante da investigação. Não basta dizer que o escopo está no FP tal como compreendido pela filosofia, pois boa parte do debate filosófico se dá justamente acerca de *como* ele deve ser compreendido. É preciso especificar o conjunto de requisitos que melhor caracteriza o FP, e esta não é uma tarefa simples. Em certa medida, todo o restante da presente investigação pode ser também compreendido como uma tentativa de realizá-la, bem como de elucidar os possíveis desdobramentos.

[32] O termo "ontologia" é algumas vezes utilizado por cientistas da computação num sentido relacionado, porém de maneira muito vaga, com o papel que o termo desempenha na filosofia. Grosso modo, trata-se da descrição dos objetos, propriedades e relações por meio dos quais se descrevem os estados computacionais possíveis em um dado sistema. Um *software* de gestão de bibliotecas, por exemplo, tem uma "ontologia" que inclui livros, usuários e relações de empréstimos entre eles.

O ponto de partida pode ser dado pela definição de um escopo: a quem o FP se mostra um problema? A resposta aqui adotada é: o FP é uma ameaça genuína para todos os que quiserem explicar a atividade cognitiva humana fazendo uso de de *representações* mentais. Neste âmbito, representações são uma forma de lidar com o fato de que a mente humana tem conteúdo intencional, isto é, seres humanos são sensíveis a significados, e não apenas a estímulos causais. Sterelny sintetiza esta suposição da seguinte forma:

> . . . na considerável medida em que nosso comportamento é flexível de modo adaptativo e informacionalmente sensível, ele precisa ser guiado por representações. Não pode haver sensibilidade informacional sem representação. Não pode haver respostas flexíveis e adaptativas ao mundo sem representação. (STERELNY, 1990, p. 21)

Esse é o escopo das teorias representacionais da intencionalidade (TRI) que buscam explicar estados intencionais em termos de estados representacionais do aparato cognitivo. Em geral, a TRI aduz à imagem de um sistema com mecanismos de entrada (*input*) e saída (*output*) de informações entre os quais dispõe-se um rico sistema representacional (STERELNY, 1990). Tal sistema precisa dispor também de mecanismos cognitivos que atualizem as representações a partir das entradas de fora e ajuste o comportamento do sistema a partir destas mesmas representações. Na maioria das teorias deste tipo, se não todas, esses mecanismos tem caráter computacional. Por isso, esse é também o escopo no qual se insere a TCC. Não raro, TRI e TCC são tomadas como indissociáveis, uma vez que o computacionalismo parece a única abordagem capaz de dar à TRI um modo de resistir a desafios clássicos como o problema do homúnculo. Contudo, não é necessário subscrever a essa tese apenas para determinar o escopo desse trabalho. Basta estabelecer que as teorias aqui discutidas serão teorias que adotam tanto a TCC quanto a TRI, e serão, daqui pra frente, denominadas *teorias representacionais da mente* (TRM).

A partir destas considerações e das discussões até aqui realizadas, estão dadas as condições para finalmente formular um (por ora um tanto vago) *requisito organizacional*. O ponto a enfatizar é o de que, na TRM, não parece possível uma solução para o FP que não envolva uma descrição detalhada sobre o modo como a informação (isto é, o conjunto de representações) do sistema é organizada e armazenada. O FP será compreendido por meio de uma síntese feita por Sheldon Chow (2013) onde ele aparece como o problema: *". . . de como um sistema cognitivo incorpora a organização informacional e permite acesso à informação certa para suas tarefas cognitivas, que parece ser requerido para performance cognitiva tal qual a dos seres humanos."* (2013, p. 15). Essa concepção apresenta um contraste com as definições mais comuns: enquanto aquelas geralmente enfatizam o aspecto inferencial do desafio, esta enfatiza seu aspecto *organizacional*. Esse aspecto, como se verá, vai constituir uma ferramenta de análise fundamental para a discussão que ainda está por vir.

Uma ilustração pode ajudar: suponha-se o objetivo de tornar acessível o conteúdo de uma biblioteca gigantesca. O sucesso desta tarefa depende tanto da elaboração de um processo de consulta ou busca (isto é, um algoritmo), quanto da elaboração de um modo de organizar os livros. Se o objetivo for localizar um livro pelo nome do autor, então o fato de os livros estarem organizados por ordem alfabética de autor é relevante para o desempenho da busca, pois será possível "saltar" seções inteiras realizando poucos testes (se a seção atual não for aquela cujos nomes de autores comecem com a primeira letra do nome do autor desejado, pode-se saltar para a próxima e só então repetir o teste). Por outro lado, se os livros estiverem organizados pela cor da capa ou estiverem aleatoriamente dispostos, será necessário realizar um teste para cada livro, tornando o processo extremamente ineficiente ou computacionalmente intratável.

Este exemplo permite também elucidar um pouco melhor o que faz do FP um desafio tão difícil. A tarefa de organizar os livros parece simples, mas tal como no caso dos sistemas de busca, isto é enganoso. É a própria biblioteca que determina os modos pelos quais uma busca pode ser realizada em função do modo pelo qual os livros estão organizados. Se for solicitado ao bibliotecário o uso de um critério de pesquisa como "livros cuja contracapa contenha uma foto anterior à 1970", ele se limitará a desejar boa sorte. Para que um tal critério pudesse se mostrar eficiente, seria preciso reorganizar completamente o acervo. Se a cognição humana for tomada como um modo de lidar com um acervo representacional, como de fato o cognitivismo o faz, então a flexibilidade e a capacidade de adaptação exibidas no comportamento humano só parecem explicáveis se a organização da informação for continuamente revista em função do contexto. O autor que melhor captou este aspecto do FP desde seus primórdios foi Janlert. Ele diz:

> (...) a julgar pelo modo como seres humanos lidam com um mundo complexo, o agente pode precisar [representar] diversas versões do mundo, cada uma com categorizações sob medida para diferentes propósitos, alternando entre elas de acordo com o que é tomado como relevante no momento"(JANLERT, 1996, p. 38)

As circunstâncias parecem afetar não apenas o critério de busca, mas também a organização da informação pesquisada, isto é, o próprio mundo do agente. Isto permite jogar alguma luz sobre o motivo pelo qual tentativas de *fixar* e tipificar a organização informacional (abordagem *cheap test*), bem como de fixar os parâmetros por meio dos quais se navega a informação (*sleeping dog*) mostraram-se infrutíferas. Todas tomaram como certo que haveria um modo *fixo* de organizar a informação, e que a flexibilidade da cognição humana repousaria sobre um algoritmo por meio do qual ele usa esta informação. Isso caracteriza uma ênfase no aspecto inferencial que, embora presente em boa parte da literatura acerca do FP, mostra-se enganadora. Não raro, é essa ênfase que faz com que alguém sinta-se justificado em negar T2 sem abrir mão da TRM, por exemplo. O problema é que um arranjo organizacional fixo permitiria

lidar com vários tipos de situação, mas não com todas as situações concretas em que um sistema cognitivo pode se encontrar no mundo real. Vem daí a aporia encontrada desde as primeiras sugestões de solução: toda tentativa mostrava-se insuficiente diante de circunstâncias que demandavam uma reorganização do modo como o mundo é representado para o agente cognitivo, e toda tentativa de suprir esta insuficiência mostrava-se *ad hoc*, afinal, ela só tinha serventia para um conjunto específico de mundos (isto é, de arranjos informacionais). Como o próprio Janlert sintetiza: *"o frame problem não tem uma única solução, existem pelo menos tantas soluções quanto existem mundos."* (1996, p. 43). Este é um modo de dizer que o FP não é constituído pelo desafio de encontrar *o* arranjo informacional por meio do qual o ser humano sedimenta as informações que usa em sua lida com o mundo, pois isto seria o equivalente a encontrar o arranjo utilizado em um único mundo, isto é, em uma única circunstância concreta específica. O verdadeiro desafio é explicar como é possível que ele consiga dispor, concomitantemente, de tantas versões de mundo diferentes, isto é, de potencialmente infinitos arranjos organizacionais, cada um sendo a contraparte de uma situação concreta com a qual ele lida. Assim, é preciso explicar não apenas como uma tal dinamicidade é possível, mas também como é possível que o aparato cognitivo do ser humano consiga navegar entre esse mar de arranjos com tanta eficiência. Neste sentido, Hayes acertou ao classificar o FP e o FPIA como problemas ontológicos.

Evidentemente, não se está a defender que a cognição humana dependa de uma literal reorganização radical e contínua de todo o conteúdo informacional a que ela tem acesso no sentido proposto. Até porque uma tal "explicação" não seria mais do que uma reformulação do FP: como determinar o critério por meio do qual o sistema detecta a organização adequada para uma determinada demanda cognitiva? Tal critério seria válido para alguns arranjos informacionais, mas não outros. O ponto é que uma solução para o FP assim compreendido parece demandar a explicação de um fenômeno equivalente e para o qual, mesmo hoje, não se tem muitas pistas.

1.6 Considerações finais

No início deste capítulo, deixou-se em aberto qual seria a possível relação entre o FPIA e o FP. Dado o que foi apresentado, é possível agora expressar esta relação de modo um pouco mais preciso: ambos os problemas são espécies de um gênero comum, que é o PGR. Isto vai de encontro à posição de autores como Shanahan (1997), para quem os dois problemas tem natureza distinta, mas também contrasta com autores como CROCKETT (1994), para quem a solução de ambos deve ser uma só. Trata-se, portanto, de uma via metodológica média: ambos são versões de um mesmo problema geral (PGR) para os quais só existem soluções específicas. A especificidade das soluções possíveis para cada um destes problemas pode ser expressa pelo conjunto de requisitos que os constituem. No caso do FPIA, por exemplo, o requi-

sito antropocêntrico não é uma necessidade, e isto faz toda diferença para o campo das soluções possíveis. Ao mesmo tempo, o requisito organizacional descreve um elemento comum, tanto ao FP quanto ao FPIA. Por meio dele, pode-se compreender de modo menos vago o motivo pelo qual as tentativas de solucionar o FP por meio de diferentes processos cognitivos oscilam entre reformulações do problema e soluções com caráter *ad hoc*: elas assumem um arranjo organizacional fixo.

O requisito organizacional permite também adotar uma postura intermediária entre o posicionamento de Fodor e o de McDermott e Samuels. Como exposto, Samuels acredita (e McDermott concordaria) que a caracterização de processos cognitivos como heurísticos permite defender a existência de formas de circunscrever os elementos relevantes (ou um conjunto suficientemente coincidente) que não são computacionalmente intratáveis. Contudo, a eficácia destas heurísticas depende do arranjo organizacional em que se inserem, isto é, do modo como o mundo é representado para ao sistema cognitivo. Como os arranjos devem ser, eles mesmos, sensíveis às circunstâncias, disto se segue que não há um único arranjo possível, mas inúmeros. É preciso, portanto, selecionar o arranjo adequado a cada situação, mas esta seleção não pode se dar por heurística, pois a eficiência deste processo seria, ele mesmo, dependente de um arranjo adequado previamente selecionado. Assim, a determinação de possíveis arranjos por meio de heurística dependeria de um arranjo de ordem superior. Para evitar um regresso sem fim, seria preciso postular, em algum momento, um arranjo e um conjuntos de operações fixas e insensíveis às circunstâncias, o que equivale a retornar à estaca zero. Deste modo, não é preciso recorrer à sugestão fodoriana de que a sensibilidade ao contexto envolve, total ou parcialmente, o acesso a propriedades globais. O requisito organizacional do FP pode cumprir este papel, além de apresentar um desafio que soa, intuitivamente, ainda mais difícil. Se propriedades globais pareciam um ponto visível num horizonte que nunca chega, o aspecto organizacional parece capaz de deixar o cognitivista numa completa escuridão. Fodor talvez concedesse este ponto sorrindo.

Uma última consideração importante diz respeito ao papel que o FP tem na ciência cognitiva e IA contemporâneas. Problemas de relevância são relativamente fáceis de caracterizar. Basta compreender que o desafio maior não é o de elaborar uma solução adequada, mas sim o de identificá-la em meio a inúmeras outras possibilidades. Explicar a resiliência de tais problemas, por outro lado, é uma tarefa bastante demandante. Sem uma familiaridade mínima com o histórico de tentativas de solução para o FP e com o tipo de dificuldade que estas tentativas enfrentaram, é grande a tentação de tratá-lo como um problema menor ou muito específico. Talvez como algo que vá desaparecer na medida em que novas tecnologias formalismos ou arquiteturas computacionais (como o conexionismo) se apresentem. É possível especular que esta visão turva contribui largamente para o aquiescimento do debate sobre o FP. Além disso, a IA contemporânea caracteriza-se antes como uma empreitada de engenharia com fins específicos e contextos bem delineados (jogar xadrez, simular

conversas sobre temas específicos etc). Mesmo uma IA capaz de dirigir um carro lida com um recorte muito específico de condições que viabilizam seu sucesso. Isto faz com que os esforços passem ao largo de onde o FP aparece: a sensibilidade às circunstâncias concretas. O FP não é, portanto, um problema real para a maior parte destas empreitadas. Estes contextos específicos dentro dos quais os sistemas são construídos acabam funcionando como um âmbito (emprestando o termo de Fodor) *informacionalmente encapsulado*, isto é, constituem micro-universos. É o caso de sistemas especialistas, como os que buscam realizar diagnósticos específicos a partir de dados do paciente. Mesmo projetos que apresentem aparente flexibilidade são fruto de micro-universos mais complexos (ou de conjuntos destes) e não de uma capacidade real de lidar com elementos inter-contextuais. Processar uma pergunta como “qual o e-mail do Beethoven?” e reconhecê-la como estapafúrdia, ou mesmo como uma piada, envolve uma capacidade de navegar entre possíveis contextos e estabelecer relações entre eles. Tanto na GOFAI quanto na IA contemporânea, a simulação desta capacidade demandaria o processamento de um volume gigantesco de informações, não apenas sobre os conceitos envolvidos, mas também sobre os possíveis contextos em que estes conceitos podem ser utilizados. É possível, claro, criar mecanismos *ad hoc* que busquem capturar este tipo de efeito. Se o objetivo for resolver um problema de engenharia, este passo é suficiente. Porém, se o objetivo for insistir no uso de abordagens cognitivistas para explicar o modo como a mente humana opera, este tipo de solução pouco ou nada tem a contribuir.

2 Pensamento e sensibilidade ao contexto

No primeiro capítulo, delineou-se o modo pelo qual um problema técnico (o FPIA) veio a se mostrar um enorme desafio na empreitada para modelar a mente e a inteligência humanas: o *frame problem* (FP). O FP foi então apresentado como uma espécie do PGR (problema geral de relevância) cujo quadro em que se apresenta é descrito pelos requisitos computacional, antropocêntrico e organizacional. Esta descrição, claro, não pretende esgotar tudo o que há para dizer sobre o FP. Neste capítulo, a discussão orbitará um possível requisito adicional denominado *requisito sentencial*. Antes de apresentá-lo, porém, é preciso dar mais alguns passos. Como visto, no caso da inteligência humana, o FP se manifesta com ênfase no tratamento da sensibilidade às circunstâncias, isto é, na sensibilidade ao contexto da atividade cognitiva. Até o momento, porém, a noção de sensibilidade ao contexto foi tratada de modo relativamente vago e intuitivo. É necessário agora refinar essa compreensão. O que significa exatamente dizer que um determinado processo cognitivo consegue tratar adequadamente de fatores contextuais?

Para responder a essa pergunta é preciso, em primeiro lugar, circunscrever o nível de análise em que ela se coloca. É possível, por exemplo, caracterizar o modo por meio do qual a sensibilidade ao contexto se dá no âmbito da linguagem e da comunicação. Da mesma forma, pode-se também descrever a sensibilidade ao contexto no âmbito da intencionalidade. É verdade que a distinção entre estes dois níveis nem sempre é clara, mas isso não impede que os dois âmbitos recebam tratativas distintas, nem que sejam explicados por teorias distintas. Esse é um ponto importante a salientar, visto que a discussão proposta nesse capítulo não se encaixa exatamente em nenhum destes campos, sendo melhor localizada num terceiro espaço: o âmbito do *veículo do pensamento*. Este nível preocupa-se com o modo pelo qual a informação é representada por mecanismos mentais. Embora autores como Fodor costumem tratar da intencionalidade e de representações mentais como constituindo um mesmo problema, estes se localizam em níveis de análise distintos.

Para esclarecer este ponto, por vezes diluído em discussões confusas, basta lembrar que a explicação de estados intencionais em termos de estados representacionais não é um dado da TRM, mas uma de suas principais teses. Na TRM, a pergunta “em função de que pode um dado pensamento ser sobre X?” é respondida por apelo a estados representacionais: o pensamento é sobre X porque o pensamento é constituído por representações de X. Assim, ao explicar “em função de que um dado estado físico pode ser uma representação de X?”, explica-se, ao mesmo tempo, os meios pelos quais um sistema cognitivo pode conter representações e como ele pode constituir

um sistema intencional.

A pergunta que se coloca, portanto, é: como pode o veículo do pensamento apresentar sensibilidade ao contexto? Normalmente, neste nível de análise, há limitações não presentes nos demais. No caso da comunicação, por exemplo, pode-se apelar a contextos discursivos ou intenções do falante, a partir dos quais o ouvinte capta o significado daquilo que foi dito. Do mesmo modo, tanto no nível da comunicação quanto no nível da intencionalidade, pode-se acomodar, mais facilmente, elementos externos que participem ou influenciem na determinação do conteúdo. Em contraste, no caso do veículo do pensamento, parece haver pouco espaço para este tipo de artifício. Como pode um determinado conteúdo ser carregado senão por meio daquilo que é explicitamente representado no veículo? Tome-se como exemplo um pensamento exprimível por "sou alto demais para sentar na cadeira". Trata-se de um conteúdo repleto de referências contextuais. Para onde aponta o termo "sou"? A qual cadeira o pensamento se refere? Trata-se de uma cadeira apontada, desenhada ou imaginada? Qual o critério utilizado para considerar-se "alto demais"? Na TRM, para que um pensamento como este possa ser expresso, todos os elementos contextuais precisam estar presentes no veículo do pensamento, na forma de uma representação. Um argumento para este fim é formulado por VICENTE (2010) da seguinte forma:

> ...se o veículo de um pensamento fosse inexplícito sobre seu conteúdo, então o ser pensante teria que interpretá-lo, adicionando informação contextual à informação que o veículo carrega. Então, ou toda informação é posicionada conjuntamente em um veículo que a carrega (que seria então apropriadamente chamado de veículo do pensamento) ou o regresso segue adiante. (2010, p. 72)

O argumento de Vicente, tem bastante força. O objetivo da TRM, convém lembrar, não é apenas oferecer um modelo da mente humana, mas também uma explicação do comportamento inteligente apresentado por seres humanos. Ao contrário de criaturas mais simples, cujo comportamento pode ser explicado em termos de relações diretas com o seu ambiente (ou via atividade cognitiva que depende de uma afinação prévia e fixa com o ambiente), a explicação do comportamento inteligente humano precisa levar em conta o modo como os seres humanos se adaptam a diferentes ambientes e conjuntos de circunstâncias. Presumivelmente, a capacidade que o ser humano tem de compreender a si mesmo como estando em uma situação ou outra é parte do que deve ser explicado por qualquer teoria da cognição. Assim, dado que a TRM opera com uma teoria computacional da cognição, e que computações só podem ser realizadas por meio de elementos explicitamente articulados, segue-se que somente aquilo que se faz presente por meio destes elementos explicitamente articulados pode fazer parte da atividade cognitiva e, consequentemente, pode fazer parte do que constitui uma explicação do comportamento inteligente. De outro modo: para que possa viabilizar o comportamento adequado, é preciso que o sistema cognitivo do agente represente a situação em que ele se encontra. Este processo de "explicitação"

é precisamente o processo de interpretação a que Vicente se refere e será daqui por diante denominamo *requisito da explicitude*. De fato, como visto no capítulo anterior, todas as tentativas de lidar com o FP (estereótipos, por exemplo) pareciam desafiadas por este requisito. Toda tentativa de lidar com conteúdo sem representá-lo, abria espaço para situações em que o sistema se mostrava insensível a variações contextuais. A dificuldade em lidar com o FP parece confirmar que, no âmbito do veículo do pensamento, nada que participe da explicação do comportamento inteligente pode ficar "em aberto".

Se este diagnóstico for correto, o defensor da TRM parece ter diante de si dois elementos fundamentalmente inconciliáveis: de um lado, a explicitude, do outro, a sensibilidade ao contexto. O FP seria então o inevitável sintoma desta tensão fundamental. Que pode ele responder diante disso? Uma primeira possibilidade é questionar frontalmente o requisito da explicitude: será mesmo o caso que a TRM demande que todo conteúdo seja representado? *Prima facie* parece, de fato, um tanto estranho supor que poderia haver algum tipo de conteúdo inexplícito em um sistema computacional. No entanto, é possível levantar um primeiro ponto importante contra Vicente a partir de CUMMINS (2010). Dentre outros exemplos, Cummins apresenta a noção de *conteúdo implícito de controle*. Trata-se de uma aparente consequência do caráter sequencial dos processos computacionais e algorítmicos.

Suponha-se um algoritmo qualquer em que uma dada operação H é executada apenas quando uma dada condição C for satisfeita. Uma vez que H seja executada, não é necessário que as condições que constituam C permaneçam explicitamente representadas para que a informação de que elas foram satisfeitas seja carregada implicitamente por H. Um exemplo um pouco mais concreto pode ajudar: tome-se um programa de xadrez que, em função do algoritmo que o constitui, só admite movimentar a rainha se os dois bispos já tiverem sido capturados. O programa pode ter um registro interno que especifica se ele está disposto ou não a mover a rainha. O valor deste registro pode ser "sim" ou "não". Seu valor inicial é sempre "não", e o algoritmo o altera para "sim" apenas na ocasião da captura do segundo bispo. Com efeito, se o valor de tal registro for "sim", isso significa que os dois bispos já terão sido capturados. Em virtude disso, será possível atribuir conteúdo "intencional" ao programa: ele agora "acredita" que já pode começar a movimentar a rainha, dado que os bispos foram ambos capturados. Contudo, em lugar algum do sistema é necessário (embora seja possível) que haja uma representação da informação "bispos fora do jogo" ou algo nestes moldes.

Um caminho para rejeitar esse exemplo é afirmar que ele não é admissível na TRM. Ele não mostra como é possível haver conteúdo implícito na TRM, mas constitui um contraponto à própria teoria, visto mostrar que ela não admite, nem acomoda, este tipo de conteúdo. No exemplo, a atribuição da crença de que os bispos foram removidos do jogo é explicada por apelo ao comportamento do sistema, e não a por meio de um conjunto de representações. Esta manobra, segue a objeção, não seria aceitável

na TRM. Contudo, pode-se replicar que não há qualquer tensão fundamental entre a TRM e a admissão de que nem todo conteúdo intencional oriunda de conteúdo representacional. O conteúdo não é atribuído ao agente a partir do seu comportamento, pois permanece sendo fruto do esquema representacional, sendo possível, inclusive, identificar e circunscrever a porção da estrutura representacional responsável por carregar implicitamente aquele conteúdo. Esta explicação do conteúdo intencional a partir de conteúdo implicitamente carregado permanece compatível com a semântica interpretacional adotada pela TRM.[1]

Na mesma linha desse primeiro exemplo, Cummins argumenta também pela existência de conteúdo implícito nas regras que o sistema cognitivo executa sem representá-las. Tal como a máquina universal de Turing demonstrou ser possível, computadores podem ser programados, isto é, podem executar conjuntos de regras diferentes daquele conjunto geral que os constitui. Os algoritmos destes programas precisam estar explicitamente representados para que o computador possa lê-los e executá-los. Contudo, este não é o caso das regras e operações formais que constituem o próprio computador. A operação de verificar se há um programa a ser carregado, ou a de checar qual a próxima instrução do algoritmo a executar, por exemplo, são apenas seguidas, e nenhuma delas está explicitamente representada. A realização destas operações é um efeito direto do veículo físico a partir do qual o computador fora construído (componentes eletrônicos, por exemplo). Nesse sentido, tais regras são ditas "embutidas" ou "incorporadas".[2] Elas fazem parte do que constitui os mecanismos cognitivos do sistema em questão, sendo portanto parcialmente responsáveis pelo comportamento apresentado por ele. Ora, se for este o caso, e se a mente for de fato um computador, segue-se que as regras formais que o constituem são incorporadas pelo agente, e não representadas. Isso significa que é possível atribuir conteúdo intencional, tal como atitudes proposicionais, a um sistema cognitivo mesmo na ausência de uma representação explícita subjacente. É o que se passa no exemplo do sistema de xadrez, que pode operar a partir da "crença" de que os bispos já foram capturados, mesmo que esta informação não esteja representada.[3] Cummins sintetiza

[1] Conforme descrito no capítulo anterior.

[2] No âmbito das ciências da computação, e em parte da literatura filosófica que orbita a GOFAI, o termo geralmente utilizado para esta característica é "*hardcoded*". Contudo, evita-se o uso deste termo aqui devido a seu sentido duplo. Ele pode ser utilizado também para referir-se a variáveis fixas explicitamente representadas dentro de um programa. Suponha-se um algoritmo que possa ser usado para buscar por uma palavra chave em arquivos de texto quaisquer. O algoritmo pode ter um limite fixo de até 100 palavras encontradas, independentemente do tamanho do texto procurado (que pode apresentar muito mais ocorrências da palavra). Este limite é também denominado *hardcoded*.

[3] Cabe esclarecer um ponto. Esta forma de conteúdo implícito pode parecer um passo na direção de alguma formade cognição situada, tal como defendido por autores como CLARK (1998) ou CHEMERO (2009). Sim e não. Sim, porque há várias formas de "situar" um sistema cognitivo que são compatíveis com a TRM (para uma visão geral de algumas possibilidades, vide ZIEMKE (2003)). Não, porque o sistema cognitivo permanece em contato com o mundo apenas por meio de representações. Considerar que este tipo de conteúdo implícito constitui uma forma de cognição situada seria uma questão meramente terminológica, portanto. Além disso, o conhecimento e aprendizado do agente precisa ser, necessariamente, explicado por meio do armazenamento de representações. Não bastasse, o conjunto de regras incorporadas é necessariamente fixo, o que significa que ele seria, ao menos em larga me-

este ponto da seguinte forma:

> . . . ainda que o sistema não represente tais regras, o fato de que ele as executa é equivalente à presença de informação proposicionalmente formulável no sistema, informação esta que não é explicitamente representada, mas é inexplícita em virtude da estrutura física sobre a qual a execução do programa se dá. (2010, p. 94)

Admitida esta hipótese, o que isto significa para o requisito da explicitude? Talvez Vicente não tivesse problema em aceitar a existência deste tipo de conteúdo implícito. O ponto central de sua tese, afinal, é estabelecer que nenhum elemento contextual pode participar da explicação do comportamento humano a menos que este seja explicitado no veículo do pensamento. Os exemplos de Cummins parecem mostrar a possibilidade de que um conteúdo implícito exerça algum papel no comportamento, mas não há clareza sobre se e como este tipo de conteúdo poderia participar do tratamento de efeitos contextuais. O importante a notar, por ora, é que o argumento de Vicente não é suficiente para estabelecer, por princípio, a impossibilidade de que este tipo de conteúdo implícito auxilie na resolução de dependência contextuais.

Para elucidar esta questão, é preciso dar um passo a mais e fornecer razões para pensar que nenhum tipo de conteúdo implícito admissível na TRM é suficiente para explicar a sensibilidade ao contexto. Isso não é difícil, visto que, de fato, os exemplos de Cummins não parecem muito promissores se aplicados a este fim. Como o conjunto de operações é fixo, não parece haver nenhuma boa razão para acreditar que este tipo de conteúdo implícito seja capaz de responder pelo tratamento adequado de todas as dependências contextuais. Assim, não há nenhuma boa razão para acreditar que esta noção de conteúdo implícito pode ajudar na solução do FP. Porém, também não parece possível descartar de pronto a tese de que conteúdos implícitos não tenham papel algum a exercer no tratamento de informação contextual. A conclusão preliminar mais razoável a partir do exposto é de que ainda não é possível medir a relação entre o requisito da explicitude e a sensibilidade ao contexto. Com efeito, para avançar nesta discussão, é preciso antes desenvolver um pouco mais os possíveis tipos de dependência contextual admissíveis no veículo do pensamento. Este será o tema da próxima sessão.

2.1 Variedades da dependência contextual no veículo do pensamento

Há espaço para dependência contextual no veículo do pensamento? Se sim, que tipo de dependência seria esta? O contexto é um fenômeno que pode se apresentar de muitos modos, a depender do escopo da investigação e da teoria em questão.

dida, inato. Com efeito, nada ligado à existência de conteúdo implícito nos termos aqui expostos vai contra as suposições mais elementares do cognitivismo ou da TRM.

No âmbito da linguagem e da comunicação, por exemplo, pode haver formas de dependência que inexistem no caso do veículo do pensamento (além de recursos para resolvê-la). Contudo, dado que os fenômenos contextuais foram amplamente estudados neste meio, o caminho aqui adotado é o de adaptar os tipos de dependência contextual já mapeados na linguagem e na comunicação. Uma autora que realizou tal mapeamento foi BIANCHI (1999). Ela enumerou três diferentes tipos de dependência, que servirão como ponto de partida. Os dois primeiros tipos são equivalentes ao que RECANATI (2007) denominou *indexicalidade em sentido estrito* e *indexicalidade em sentido amplo*. O terceiro tipo é denominado por Bianchi *meta-dependência contextual*. Como se verá, efeitos deste terceiro tipo constituem o maior desafio para a TRM e campo fértil para o FP. Segue-se agora um pequeno desenvolvimento de cada um deles.

Os *indexicais em sentido estrito* são caracterizados pela existência de regras pré-definidas que apontam para uma ou mais variáveis contextuais, também definidas de modo fixo. Uma ocorrência de “aqui”, por exemplo, invariavelmente captura a variável contextual “lugar em que o falante se encontra”. Na mesma linha, a ocorrência de “eu” aponta sempre para a variável “falante”, e assim por diante. No caso da linguagem, estas regras que conectam expressões e variáveis contextuais podem ser explicadas de vários modos, mas em geral, a ferramenta usada é a sedimentação de convenções. Estas podem ser fixadas durante o processo de maturação do agente, o que inclui sua iniciação às convenções de sua comunidade linguística. Assim, aquilo que é dito pelo uso do indexical em uma enunciação pode ser dado pelo significado do termo. O conteúdo da variável contextual pode depender das circunstâncias, mas a regra por meio da qual esta variável é capturada não varia.

No caso do veículo do pensamento, esse tipo de dependência contextual pode ser facilmente acomodada a partir de referências indiretas. Os indexicais podem ser símbolos que apontam para regras fixas. Estas, por sua vez, apontam para outras representações, ou seja, apontam para o conteúdo das variáveis contextuais. Com efeito, a ocorrência de uma representação X exprimível por “aqui” apontará sempre para as representações que descrevem o local em que o agente se encontra. Assim como aquilo que é dito em uma enunciação pode ser dado pelo significado do indexical expresso, aquilo que é pensado pode ser dado pelo conteúdo para o qual a ocorrência de um indexical indiretamente aponta. Este modo de compreender indexicais no veículo do pensamento é perfeitamente compatível com o requisito da explicitude, e não constitui um desafio para a TRM. De modo especialmente importante, contudo, alguns casos deste tipo de dependência podem ser resolvidas por apelo a conteúdo implicitamente carregado. Um indexical como “eu”, por exemplo, poderia ser sempre resolvido por processos total ou parcialmente incorporados. Evidentemente, defender que é, de fato, o caso que este ou outros indexicais são resolvidos desta forma, é uma tese que demandaria muito mais esforço argumentativo para ser estabelecida. Não foi possível localizar quem defenda tal tese, e não é o que se pretende fazer aqui. O objetivo é

apenas apontar que não parece haver qualquer barreira fundamental a impedir que, pelo menos alguns indexicais estritos sejam resolvidos a partir de conteúdo implícito.

Por sua vez, os *indexicais amplos* caracterizam-se pela ausência de uma regra fixa por meio da qual a variável contextual pode ser determinada. Um exemplo deste tipo de dependência é a que se apresenta nos *demonstrativos*. Termos como “essa”, “aquela” ou “ela” não possuem regras fixas nem variáveis contextuais fixas atreladas ao seu significado linguístico. Nestes casos, os elementos linguísticos podem apenas subdeterminar o significado, que precisa então ser determinado via processos adicionais. Na linguagem, isso geralmente envolve a introdução de um processo pragmático que tenta determinar o significado correto a partir de uma dada enunciação. Neste processo, são levados em conta elementos extra-linguísticos, geralmente oriundos do caráter cooperativo da atividade comunicativa, tais como a intenção do falante: se ele aponta o dedo para um local e diz “ali”, é provável que o local apontado participe do significado da expressão.[4] Outro exemplo de indexical amplo são construções como “o livro do João”.[5] Nestes casos, o significado linguístico apenas sub-determina as condições de verdade da sentença: embora a expressão aponte para uma relação entre João e o livro, não é possível extrair dela a natureza desta relação. O referente pode ser o livro que João escreveu, emprestou ou perdeu, sem que nenhuma pista sobre qual é o caso seja dada pelo significado linguístico.

No caso da comunicação, esta relação pode ser determinada a partir das intenções do falante. Enquanto atividade cooperativa, a comunicação circunscreve o escopo de possíveis relações (posse, autoria, etc.), dando ao agente um norte a ser seguido. ANDLER (2003) formulou uma analogia que pode ser útil na apresentação deste ponto: o processo de determinação das intenções do falante é análogo ao modo pelo qual um aluno responde a uma questão de prova. Ele sabe que uma pergunta adequada não deve exigir dele mais do que um certo corpo de conhecimento previamente delimitado pelas informações contidas na própria pergunta e por um conhecimento das intenções do Professor de referenciar apenas temas previamente circunscritos. Da mesma forma, é possível partir do princípio de que o falante disponibiliza todas as pistas necessárias para a determinação daquilo que é dito, visto que ele tem a intenção de ser compreendido. Assim, num cenário como este, mesmo um conjunto limitado de estratégias cognitivas pode determinar, com grande confiabilidade, qual a natureza da relação entre João e o livro.

Pode o veículo do pensamento conter indexicais em sentido amplo? Vicente (2010) notou que este tipo de ocorrência também pode ser explicada por apelo a regras fixas, tal como as que regem o uso de indexicais não amplos. Estas regras pode descrever conexões explícitas entre o conteúdo representado sendo processado e o conteúdo da percepção ou da memória. Assim, um pensamento exprimível por “o livro

---

4 Discussões acerca destes mecanismos são frequentemente pautadas pela teoria griceana da comunicação (GRICE, 1991).

5 Este exemplo é utilizado por Recanati (2007).

do joão" pode ter a natureza da relação determinada a partir de uma conexão com um elemento da memória que contenha a informação de que o livro em questão é aquele que João emprestara. Da mesma forma, um pensamento expresso por "aquele livro" pode ter o referente determinado pela conexão a um conteúdo presente nos mecanismos perceptuais. Com efeito, indexicais amplos não apresentam qualquer tensão com o requisito da explicitude e podem ser facilmente acomodados pela TRM, visto que também podem ser tratados por referências indiretas. É verdade, contudo, que provavelmente nenhum caso de indexical amplo pode ser explicado por apelo a conteúdo implícito. Indexicais amplos estão geralmente ligados a fatores culturais ou sociais que se sedimentaram no decorrer do processo de maturação do sistema cognitivo. Como a TRM cognitivista supõe que o conjunto de regras incorporadas são fixas, defender a tese que um dado indexical amplo pode ser resolvido por apelo a conteúdo implícito carregado por tais estruturas não parece um caminho frutífero. A conclusão preliminar, portanto, é que indexicais amplos já constituem razão suficiente para negar que os exemplos de Cummins sejam um problema para a tese de Vicente.

Há, contudo, um terceiro tipo de dependência que Bianchi denominou *meta-dependência contextual*. Nesse caso, o modo como o significado de uma expressão depende do contexto é, ele mesmo, dependente de contexto, que pode neutralizar ou enfatizar a necessidade de articulação de uma ou mais variáveis contextuais. Tome-se como exemplo a sentença abaixo:

(4) O gato está sobre o tapete.[6]

Em condições típicas, a presença de um campo gravitacional não está entre as variáveis contextuais disponíveis para articulação, de modo que possa participar das condições de verdade de (4). Contudo, alterações no contexto a partir do qual tal *input* é processado podem fazer com que esta articulação seja necessária. Será o caso se o pano de fundo contra o qual a sentença for avaliada alterar-se de "planeta terra" para "espaço interestelar", por exemplo. Se nos tipos de dependência anteriores o contexto fornecia o conteúdo para as pontas que o *input* a ser processado deixava soltas, ele agora é responsável pela informação de quantas e quais são as pontas soltas que devem ser resolvidas. Assim, o que caracteriza esta forma de dependência é uma espécie de holismo: sem uma noção prévia e completa do contexto, a tarefa cognitiva não pode ser concluída adequadamente, pois é o contexto que diz quais serão as variáveis contextuais a buscar em cada situação concreta.

Esse é um tipo de dependência contextual especialmente desafiadora para a TRM. Se o mapeamento e resolução das variáveis contextuais depende de um prévio acesso a todo o contexto, disto se segue que o contexto precisa estar inteiramente representado no veículo do pensamento. O problema, como argumenta AN-

[6] O exemplo é inspirado na argumentação de Searle (1978), que primeiro fez uso dele.

DLER (2000a), é que contextos são insaturáveis.[7] Qualquer elemento do mundo, presente ou ausente, pode vir a demandar a articulação de alguma variável contextual. O que exatamente constitui o pano de fundo adequado contra o qual os estímulos oriundos de um jantar em um restaurante devem ser processados? Deve a ausência de colegas de trabalho ser levada em conta na hora de escolher um prato? Em que medida é adequado que a temperatura do ambiente influencie na escolha da bebida? Como já discutido, a necessidade de representar e checar esse grande número de possibilidades e recortar aquelas que são relevantes para a situação concreta é o que fez com o que FP emergisse na TRM. Isso indica que a meta-dependência contextual está em tensão com o requisito da explicitude: ambos são tomados como necessários para os objetivos da TRM, mas ela não consegue satisfazer os dois ao mesmo tempo.

Para Bianchi, as duas primeiras formas de dependência contextual (indexicais estritos e indexicais amplos) lidam com o que ela denomina *contexto proximal*. Em contraste, meta-dependências contextuais lidam com o chamado *contexto distal*. Essas noções de contexto também possuem significados específicos no caso da comunicação e da linguagem, espaço em que Bianchi trabalha. Contudo, eles podem ser igualmente adaptados de modo a figurar na presente discussão.

O contexto proximal é análogo a uma determinada situação em que o agente se encontra. Esta situação funciona como pano de fundo contra o qual sua atividade cognitiva se dá. Nesse sentido, tal atividade pode ser dita *local* a um dado contexto proximal. O contexto distal, por sua vez, é muito mais amplo. É a partir dele que o agente se reconhece como inserido em um contexto proximal ou outro, isto é, como estando em uma situação de um tipo ou outra. Se a atividade cognitiva se dá tendo um contexto proximal como pano de fundo, este, por sua vez, só pode ser estabelecido contra um pano de fundo mais amplo, que é o contexto distal. Isso é o que permite ao agente reconhecer a necessidade de articular ou não uma determinada variável contextual.

Suponha-se uma situação em que dois indivíduos queiram realizar uma vídeo-conferência e pretendam agendar um determinado horário para realizá-la. Deve o fuso-horário ser levado em conta neste agendamento? Caso ambos se encontrem em uma mesma cidade, o fuso-horário não aparece como uma variável contextual relevante, mas se eles estiverem em países diferentes, ela figurará como necessária. Enquanto o pano de fundo contra o qual o primeiro caso é avaliado pode ser constituído pela cidade, ou mesmo pelo país, o pano de fundo contra o qual o segundo caso é avaliado envolve, talvez, todo o continente ou todo o planeta terra. É o contexto distal que permite ao agente reconhecer a si mesmo como estando em uma situação na qual o pano de fundo adequado é um, e não outro. É também o contexto distal que permite ao agente reconhecer-se no contexto de uma atividade comunicativa. Intuitivamente,

[7] Convém notar que a caracterização de contextos como insaturáveis, por parte de Andler, tem uma conexão muito próxima com o tipo de dificuldade que Dreyfus apontava em suas críticas ao uso de estereótipos para lidar com o FP, apresentadas no primeiro capítulo.

parece possível concebê-lo como um “contexto proximal definitivo” ou como o conjunto de todos os contextos proximais. Nessa linha, compreender a si mesmo como estando inserido em um dado contexto proximal seria equivalente a selecionar o contexto proximal adequado dentre todos os que constituem tal conjunto. No entanto, fosse este o caso, o contexto distal não seria mais do que um contexto proximal de segunda ordem. O agente só conseguiria lidar com este contexto de segunda ordem pressupondo um contexto ainda mais amplo a partir do qual poderia resolver meta-dependências contextuais. A insistência nessa linha gera, portanto, um regresso potencialmente infinito de contextos. Mas se este é o caso, de que modo o contexto distal consegue evitar esse regresso?

O que distingue o contexto distal não é uma maior quantidade de fatos que ele abarca, mas sim aquilo em função de que uma dependência contextual se mostra saliente. Enquanto no contexto proximal os efeitos se dão em virtude do seu conteúdo, atuando como uma espécie de repositório em que o agente “pinça” a informação necessária, os efeitos do contexto distal se dão em virtude de sua *estrutura*, isto é, em virtude do modo como o próprio mundo do agente se organiza. Dizer que o efeito se dá em função da própria estrutura é dizer que ele não se dá em função de seus elementos mas do seu arranjo global, caracterizando assim o caráter holístico das meta-dependências contextuais. Não há nenhum sentido óbvio em que um elemento possa ser identificado, no interior de um contexto distal, como sendo o “elemento responsável” em função do qual uma situação se mostra como tal. Tome-se (4) novamente como exemplo. Não há um aspecto particular da situação concreta que possa ser apontado como aquele em função de que se torna relevante a variável contextual *campo gravitacional*. Caso se tente apontar para um aspecto qualquer (“o gato está em uma base lunar”), imediatamente vem à tona a questão: em função de que este aspecto é o aspecto relevante? Afinal, é possível pensar em cenários alternativos em que gatos estejam em bases lunares e que, mesmo assim, o campo gravitacional não seja uma variável contextual relevante: a base lunar em questão pode ser gigantesca, habitada por milhares de pessoas, algumas das quais nasceram aqui, e para quem a gravidade é uma questão tão raramente articulada quanto para os fusos-horários no caso de um terráqueo. Tem-se assim a perspectiva de um regresso infinito. A própria tentativa de “destacar” uma característica particular da estrutura do contexto distal só faz sentido tendo um contexto distal como pano de fundo. Em síntese, a depender da estrutura do contexto distal em que a atividade cognitiva do agente se insere, alguns elementos podem lhe aparecer como mais ou como menos salientes.

Mas como tratar do papel dessa estrutura no âmbito de uma teoria da cognição? O que a constitui? A resposta varia em função do ferramental conceitual que cada teoria tem à mão. No caso da TRM, pautada pelo requisito da explicitude, trata-se de uma estrutura informacional. Aquilo que aparece ao agente como uma variável contextual a ser resolvida (na forma de um indexical amplo, por exemplo) é fruto do modo como a informação está arranjada. Há, portanto, uma evidente conexão entre o

contexto distal e o aspecto organizacional do FP. Este último é sintoma da falha em explicar como é possível que um sistema cognitivo consiga lidar adequadamente com o contexto distal.

No âmbito da TRM, as tentativas de atender, ao mesmo tempo, o requisito da explicitude, e a sensibilidade ao contexto distal como aqui exposto, caracterizam a estratégia que, daqui por diante, será denominada *eliminativismo contextual*.[8] Trata-se, essencialmente, da tentativa de lidar com efeitos estruturais típicos do contexto distal por meio da mera disponibilização de informações. O nome "eliminativismo" sugere o principal efeito da estratégia: em vez de tratar do contexto distal, este tem seus efeitos eliminados.

Isso fica mais claro se a discussão do primeiro capítulo acerca de estereótipos for rapidamente retomada. Os estereótipos das situações em que o agente pode se encontrar (estar em um restaurante, estar em um hospital, etc.) são formas de representar o conteúdo do contexto proximal. Como se viu, isto gera uma espécie de gangorra entre estereótipos super determinados (que seriam inúteis para casos reais) e subdeterminados (que deixam o sistema cognitivo sujeito ao FP). A origem dessa gangorra pode agora ser exposta de modo mais direto: as representações dos estereótipos ignoravam completamente os efeitos do contexto distal. Um estereótipo com informações excessivamente detalhadas constitui uma tentativa de eliminar meta-dependências contextuais, substituindo-as por conjuntos pré-definidos de valores possíveis e conjuntos de regras para escolher entre eles. Trata-se de uma tentativa de antever todas as circunstâncias com as quais o agente pode vir a se deparar no decorrer de sua vida. Porém, conforme discutido no primeiro capítulo, este é um esforço tão hercúleo quanto inútil, pois além de restringir excessivamente a aplicabilidade de cada estereótipo de maneira quase esdrúxula (gerando, por exemplo, um estereótipo para uma situação de *jantar-em-que-uma-mosca-está-a-incomodar-pelo-lado-direito-justamente-num-primeiro-encontro-com-alguém*), o detalhamento continua insuficiente, dado que contextos são insaturáveis. Esta continua sendo uma tentativa de "extrair" do contexto distal o elemento informacional em função de que uma situação se constitui, e como visto no caso de (4), resulta daí um regresso. Nos casos em que se usam estereótipos com grau de detalhamento baixo ou moderado, o contexto distal é eliminado na medida em que se tenta fazer uso de processos heurísticos ou dedutivos para tentar detectar as meta-dependências. Como amplamente discutido no capítulo anterior, estes métodos não são sensíveis à estrutura do contexto distal. Neles, o contexto é tomado como sendo tudo aquilo que for derivável das premissas a partir das quais o processo cognitivo se dá. Esse espaço de potencialmente infinitas possibilidades é vasculhado como se fosse um gigantesco contexto proximal, fazendo emergir o FP.

Outra forma de expressar o mesmo ponto é dizer que, no eliminativismo con-

[8] O uso termo é inspirado em Andler (1993, 2000b).

textual, as meta-dependências contextuais são tratadas como se fossem indexicais amplos cuja solução deve ser encontrada "vasculhando" um volume gigantesco de informações. Contudo, como o exemplo da atividade comunicativa sugere, indexicais amplos só podem ser tratados à luz do contexto distal. É o contexto distal que estabelece a situação presente como uma atividade cooperativa na qual é possível fazer uso de critérios que lhe permitem circunscrever a informação relevante, de acordo com a analogia professor-aluno de Andler.[9] Isto pode ser também notado em outros casos mais simples. Como distinguir um conjunto de movimentos aleatórios de um gesto? Como distinguir um conjunto de gestos não relacionados de uma ação proposital? Sem a sensibilidade ao contexto distal, o eliminativismo contextual pode tornar-se vítima do FP mesmo nestes casos mais elementares.

A partir do que fora exposto, é possível lançar novo olhar sobre as primeiras tentativas de lidar com o FP no âmbito da TRM, bem como sobre o motivo de suas falhas. Como visto, o campo de trabalho das primeiras tentativas de resolver ou evitar o problema orbita o sistema formal adotado. Esta estratégia abarca três táticas. A primeira busca criar modelos que permitem o registro das informações de um modo que explique como elas podem estar prontamente disponíveis nas circunstâncias adequadas. Este é o caso das abordagens estereotípicas. A segunda tática, normalmente usada em conjunto com a primeira, é a de realizar modificações na constituição do próprio sistema formal utilizado para realizar os modelos. A lógica não monotônica de McDermott e o *Unless* de Sandewall são exemplos. A terceira tática é a de tentar modelar diretamente os processos cognitivos. Dentre os exemplos estão o GPS (*general problem solver*) que fazia uso de um processo essencialmente heurístico, o método de prova de teoremas utilizado pelo cálculo de situação, e abordagens mistas como o STRIPS. Como as táticas não são mutuamente exclusivas, cada tentativa concreta podia ou não fazer uso de diferentes combinações delas. Assim, a crença subjacente era a de que a solução do FP viria fundamentalmente da busca pelo formalismo adequado, pela descrição do processo adequado e pela articulação do modelo adequado dentro deste formalismo. A estratégia fazia sentido, afinal a busca por formalismos e modelos adequados era o que constituía a própria empreitada da GOFAI. Nessa perspectiva, o FP nada mais é do que um dos muitos desafios de uma ampla e árdua empreitada científica, não havendo razão para imaginar que ele teria condições de questionar os próprios fundamentos cognitivistas.

Contudo, cada fracasso fornecia um novo exemplo que reforçava a tese de

[9] Há um ponto que pode gerar confusão e que demanda um esclarecimento preventivo: ao dizer que o contexto distal estabelece os contextos proximais, não se está a pressupor que o mecanismo pelo qual isso se dá é sempre o de processamento de informação por meio de um processador central. Módulos especializados, tais como os de FODOR (1983), bem como outros mecanismos evolutivamente sedimentados, desde que compatíveis com a TRM, podem ser total ou parcialmente responsáveis por estabelecer um contexto proximal em diversos casos. Nesse sentido, estes mecanismos fariam parte do que constitui a sensibilidade ao contexto distal. O que se está sugerir é que, embora esta seja uma solução plausível para alguns contextos proximais, ela é insuficiente para tratar da potencialmente infinita variedade de contextos com os quais a inteligência humana lida.

Fodor: insistir na busca de uma solução nestes termos gerará apenas reformulações do problema, e não soluções. Para Fodor, a conclusão é desoladora: não existe a possibilidade de uma teoria cognitiva que abarque processos globais (tais como os que lidam com o contexto distal), e a razão e inteligência humanas são exemplos desse tipo de processo. Fodor conclui então que não é possível explicá-las (1983). Haveria, afinal, uma tensão fundamental entre a necessidade de representações explícitas e a necessidade de lidar com o contexto distal. Segue-se que o FP é um problema intransponível para o eliminativista contextual e um argumento fatal para toda teoria computacional da cognição humana.

Uma conclusão tão incômoda não pode ser facilmente aceita, e de fato nunca foi. Com a estagnação resultante das primeiras tentativas de solução, buscou-se expandir o escopo de estratégias. Alguns autores começaram a suspeitar que aquilo que é intransponível ao eliminativismo contextual, não necessariamente é intransponível para qualquer teoria computacional da cognição. Para que a explicitude possa ser utilizada como argumento contra a TRM, seria preciso antes mostrar que não é possível haver TRM sem explicitude. Isso se potencializou pela percepção de que todos os formalismos utilizados tinham um elemento comum: a estrutura sentencial, isto é o fato de todos pertencerem a um mesmo gênero representacional. Surgiu então o diagnóstico segundo o qual o problema está na tese de que o veículo do pensamento tem estrutura linguística, tal como defendido por Fodor (1980, 2010) e que se denominará, daqui por diante, *sentencialismo*. Mas o que havia, afinal, de tão problemático na estrutura sentencial? Segundo esse diagnóstico, o requisito da explicitude não é essencial a todo gênero representacional, mas sim uma característica particular do gênero sentencial. Mais especificamente, da sua capacidade de separar claramente aquilo que é implícito daquilo que é explícito. A tese pode ser sintetizada da seguinte forma:

Tese 4 (T4): O FP é fruto do uso de representações sentenciais e de sua capacidade de distinguir claramente entre o que é explícito e o que é implícito.

Nos termos utilizados no capítulo anterior, é possível expressar a mesma tese dizendo que o FP é caracterizado por um *requisito sentencial*. A pergunta que emerge desse diagnóstico é: que tipo de representação poderia então evitar o FP? Nas próximas sessões, serão analisadas duas possibilidades: representações distribuídas e representações pictoriais. Há várias razões para considerar que estes diferentes gêneros representacionais são promissores no caminho para uma solução ao FP e no tratamento do contexto distal. No entanto, será dada ênfase naquele que costuma ser o mais comum deles: ambas são, alega-se, capazes de representar informações de modo intrínseco. Ambas permitem borrar as linhas que separam de modo claro aquilo que é considerado explícito e aquilo que é considerado implícito no caso das representações sentenciais. Esta é uma afirmação que será melhor esclarecida por meio da análise detalhada de cada um destes casos.

2.2 Representações distribuídas

O que são representações distribuídas e como se distinguem? É impossível falar nelas sem antes falar daquele que parece ser seu *habitat* natural: a abordagem *conexionista* da cognição. O motivo: conforme argumenta Cummins (1991), é um erro abordar o papel e a natureza das representações mentais de modo geral. Em particular, isto significa que a pergunta "em função de que um estado X representa Y?" não possui resposta neutra em relação à teoria cognitiva no interior da qual a questão se coloca. O conexionismo permite a adoção de teorias semânticas não disponíveis, ao menos a princípio, para teorias computacionais clássicas, como é o caso da TRM cognitivista.[10] Por isso, para que se possa apreciar adequadamente o potencial deste gênero representacional, é pertinente dedicar um tempo a uma compreensão adequada de o que é uma teoria conexionista da cognição. Evidentemente, não há como realizar uma introdução abrangente. O foco estará nas características julgadas pertinentes ao escopo desta discussão.

O conexionismo é uma abordagem quase tão antiga quanto a computação clássica, porém, seu potencial só foi suficientemente reconhecido a partir do trabalho de RUMELHART et al. (1987) e MCCLELLAND et al. (1987). Sua estrutura basilar é a de um conjunto de elementos interligados em rede. Cada um destes elementos "primitivos" é capaz de receber informações, realizar operações com ela, e disparar o resultado para outras unidades com as quais esteja conectado na rede. Dentre as implementações mais comuns estão as chamadas *redes neurais*, em que cada um destes elementos basilares é um modelo matemático idealizado de um neurônio cerebral. Nesses casos, a atividade de uma rede neural é inspirada na atividade neuronal cerebral. É especialmente importante notar o uso do termo "inspirado", visto que não se trata de um modelo exaustivo, nem de neurônios, nem do modo como eles se comunicam. Pelo contrário, a semelhança com neurônios é ainda menor do que parece. Ruído informacional e fenômenos oriundos do tempo que a transmissão da informação leva para trafegar, são aspectos deixados de lado no modelo, ainda que estejam presentes nos neurônios reais, podendo, inclusive, ser relevantes para o modo como o cérebro efetivamente trabalha. Assim, redes conexionistas não necessariamente constituem tentativas de modelar uma estrutura cerebral tal como ela se mostra, ainda que muitos autores, como CHURCHLAND (1989), defendam que isto seja possível.

Como compreender o que é ou que papel realiza uma unidade básica, isto é, um nodo? Trata-se, essencialmente, de um modelo matemático. Um nodo é um elemento que recebe um número, faz algum cálculo com ele, e emite um outro número,

[10] As expressões "teorias computacionais clássicas", "computação clássica" ou "computacionalismo clássico" serão utilizadas em referência a arquiteturas computacionais não conexionistas, tal como a Máquina de Turing. A tese de que o conexionismo constitui uma arquitetura distinta, mas capaz de realizar computações do mesmo tipo será tomada como ponto de partida de toda a exposição. Como argumenta PICCININI (2008), é possível que existam abordagens conexionistas cuja atividade não seja computacional, mas este não é o caso de nenhuma das abordagens que serão aqui discutidas.

ou seja, ele realiza uma operação matemática sobre a informação numérica que recebe. O cálculo realizado é análogo a uma tomada de decisão. A depender do modelo matemático utilizado, os possíveis resultados são bastante variáveis. No caso do *perceptron*, que está entre os primeiros modelos formulados, o resultado é binário, isto é, pode ser 1 ou 0 (ROSENBLATT, 1958). Já no caso de modelos pautados pela função sigmóide, mais comum contemporaneamente, a saída é contínua e pode variar entre qualquer valor possível de 0 a 1.

Uma maneira didática (ainda que arriscada) de facilitar a compreensão do papel destas funções matemáticas atreladas aos nodos de uma rede, é apresentá-las como se fossem semanticamente permeadas.[11] Suponha-se que a rede conexionista esteja realizando um processo cognitivo decisório: deve o agente ir à aula hoje? Os nodos são então estruturados em três camadas: a camada de *input* (que recebe os componentes da pergunta), uma camada oculta (utilizada apenas como instrumento de processamento da informação) e uma camada de *output* que disponibiliza o resultado final, isto é, a decisão tomada. Para que uma rede conexionista seja capaz de lidar com este caso, ela precisa ter sido previamente capacitada (treinada) para tal. Uma vez treinada, ela espera que cada um dos nodos da camada de *input* corresponda a um critério utilizado no processo cognitivo. Um nodo `a` pode receber um valor que identifique se está a chover ou não. Um nodo `b` pode receber um valor que identifique se a aula é dada pelo professor ou pelo estagiário. Um nodo `c` pode receber um valor que identifique o grau de interesse pelo tema da aula, e assim por diante.

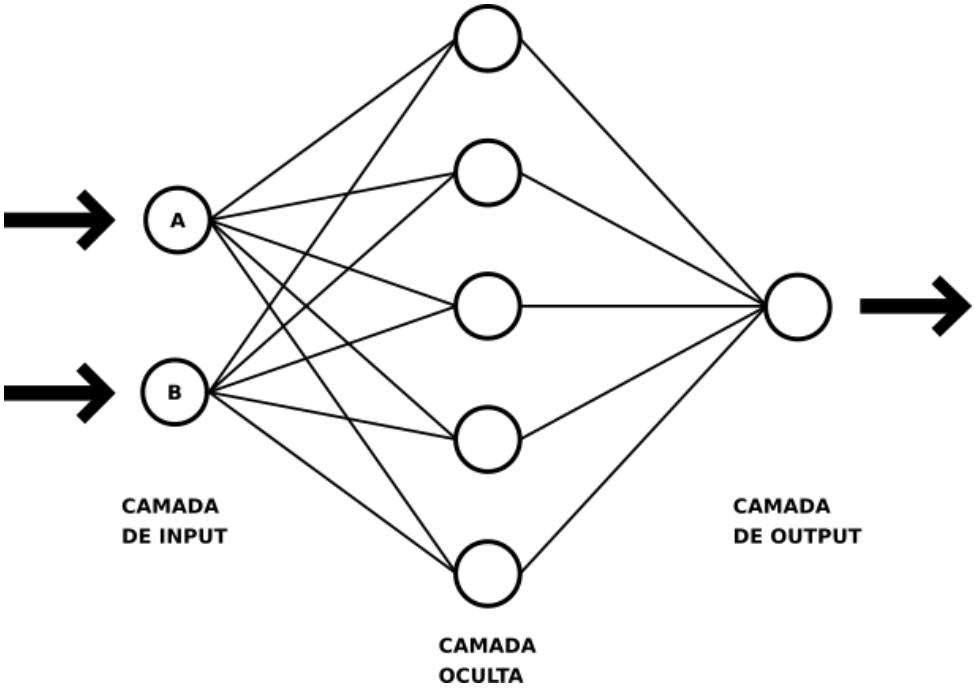

Figura 3 – Estrutura básica de uma rede conexionista *feedforward*

[11] O risco deve-se ao fato de que nodos não são semanticamente carregados. Esta é apenas uma estratégia de apresentação que tenta abrir mão da matemática pesada envolvida. Como se verá mais adiante, redes conexionistas suportam uma pluralidade de teorias semânticas. Estas é que dirão se é possível atribuir conteúdo a um nodo tomado de forma isolada ou não.

O passo seguinte é compreender qual a operação realizada com estes *inputs*. Ela varia conforme os detalhes do modelo matemático mas, de forma geral, pode-se imaginar cada nodo como uma entidade que contém dois números: um é denominado *peso*, e o outro é denominado *viés*. O peso do nodo pode ser tomado como identificando a importância que o critério a ele atrelado tem, isto é, o quanto ele "pesa" na decisão. O nodo `a` pode ter um peso maior que o nodo `b`, por exemplo. Vale notar que a diferença no peso pode ser extremamente nuançada, visto ser admissível qualquer valor entre 0 e 1. Com efeito, um *input* numericamente alto que indique (por exemplo) uma chuva fortíssima pode, mesmo assim, resultar em um *output* baixo por parte daquele nodo, pois ele atribui um baixo peso à chuva, indicando que chuvas não costumam pesar na decisão.

Além do peso, existe também um segundo valor, denominado *viés*. Sua função é essencialmente a mesma do peso, sendo usado para tratar situações de modo ainda mais nuançado. Pode-se considerá-lo indicativo de uma tendência ou sensibilidade que o sistema cognitivo tem a um certo *input*. Se, por exemplo, ele não gostar de estudar ou aprender, ainda que o peso de um nodo seja moderado ou alto, ele pode ter um viés que reduza significativamente o *input*, fazendo com que seja necessário um valor muito alto para que o sistema "vença" a tendência de não ir à aula por desinteresse.

Pode-se sintetizar, então, que uma rede conexionista é uma rede de nodos, cada qual com seu peso e viés. A topologia da rede, isto é, o modo como tais nodos são estruturados, também pode variar. É a estrutura que indica qual nodo recebe o *input* de quais outros nodos, bem como quais nodos receberão seus *outputs*. A estrutura mais simples (e comum) é denominada *feedforward*. Trata-se de uma estrutura em que os *inputs* "trafegam" da esquerda para a direita. Assim, a própria rede atua como um grande e complexo nodo, que recebe *inputs*, realiza um processamento com eles, e disponibiliza *outputs* na saída. Redes *feedforward* são especialmente eficazes num aspecto que é considerado um dos grandes pontos fracos do computacionalismo clássico: a análise de padrões, tais como imagens, letras escritas à mão, faces etc. No caso do computacionalismo, seria preciso fazer uma análise sucinta e decompor a face (ou o que quer que esteja sendo analisado) em diferentes subcomponentes (olhos, nariz, boca etc.) bem como mapear possíveis relações entre eles (nariz sempre acima da boca, olhos abaixo da testa etc.) No caso das redes conexionistas, tal análise não é necessária.

Um modo de compreender como isto se viabiliza é comparando os estímulos que o sistema visual do agente cognitivo recebe, com uma imagem de um computador. Imagens digitais são essencialmente matrizes bidimensionais em que cada posição acomoda um *pixel*. Quanto mais *pixels*, maior a resolução, isto é, maior a qualidade da imagem. Cada *pixel* é constituído por um valor atrelado a um matiz de cor e corresponde a um nodo na camada de entrada de uma rede conexionista. Assim, uma rede que consiga analisar uma imagem com 840 pixels (35x24) precisará contar com

840 neurônios de entrada. A imagem é então dividida, e cada pixel é introduzido como *input* de um nodo na camada de entrada. O *input* de cada nodo é, neste caso, um número que identifica um matiz de cor, conforme uma convenção qualquer previamente estabelecida.

Como visto, cada nodo possui um peso utilizado para tomar uma decisão. Nesse caso, o peso é uma medida da probabilidade de que o matiz de cor recebido faça parte do rosto que se está a tentar reconhecer. Como as imagens, em geral, possuem figura e fundo (como em fotos típicas, o rosto se destaca sobre um cenário de fundo), os primeiros nodos esperam encontrar parte desse plano de fundo. Assim, seu peso tende a ser menor. No caso dos nodos que representam posições mais próximas de onde se esperaria um contraste entre a cor do rosto e a cor do plano de fundo, coloca-se um peso maior. Se os nodos fossem visualmente arranjados de modo similar à matriz de matizes que constitui a imagem, o contraste entre os pesos "contornaria" o espaço da imagem em que se espera encontrar o padrão do rosto a reconhecer. O mesmo se daria para contornos "internos" ao rosto, tal como nariz, olhos, boca etc. Nesse sentido, pode-se dizer que os pesos representam o *conhecimento* que o sistema cognitivo possui, de como reconhecer aquele rosto.[12]

Esse processo de reconhecimento facial pressupõe, claro, que a rede tenha sido calibrada, isto é, que os pesos e vieses de cada nodo tenham sido previamente determinados de modo a coincidir com a estrutura do padrão a analisar. Disto emerge a questão: como a determinação destes pesos é realizada, considerando que uma rede conexionista típica pode ter milhares deles? Seria um trabalho hercúleo determinar manualmente todos estes pesos, principalmente considerando que o peso de um nodo precisa ser calibrado de modo sensível aos pesos dos demais nodos, isto é, o fato de um nodo ter um peso X precisa ser levado em conta na determinação de todos os demais. Não por acaso, a criação de redes conexionistas é realizada de modo semiautomático, a partir de algoritmos. Como estes algoritmos objetivam "treinar" a rede para que ela consiga reconhecer um determinado padrão desejado, eles são chamados de *algoritmos de aprendizado* ou *algoritmos de treinamento*.

Treinar uma rede é encontrar uma combinação de pesos (e vieses) tal que possa permitir o reconhecimento de um padrão desejado. O algoritmo mais simples (e ineficaz) possível seria um que testa todas as combinações possíveis. Evidentemente, tal abordagem não seria funcional: não só há um grande número de nodos, como o peso atribuível a cada nodo pode variar de modo potencialmente infinito, visto que todos os números entre 0 e 1 são admissíveis. Por isso, os algoritmos de treinamento fazem uso de bases de dados contendo os exemplares que constituirão as referências do padrão que a rede deve aprender a reconhecer (um conjunto de fotos de faces conhecidas, por exemplo).

---

[12] No âmbito da GOFAI, convencionou-se utilizar o termo "conhecimento" para referir-se ao conjunto de representações que participa do sistema cognitivo de um agente. O termo não deve ser tomado no mesmo sentido em que figura nas investigações filosóficas em epistemologia.

Há uma grande variedade de algoritmos de aprendizado que podem ser utilizados. Os mais comuns são variações do *back propagation*. Nele, os pesos e vieses são inicialmente configurados para valores aleatórios e a primeira imagem da base de dados é entregue como *input* à rede. Ela processa a informação e gera um resultado, que o algoritmo pode determinar como dentro ou fora do esperado. Como o algoritmo "conhece" o *output* desejado para cada *input* (dado que ele tem acesso à base de dados usada no treinamento, ele consegue identificar um erro quando, por exemplo, a rede é alimentada pela imagem de um cachimbo e reconhece ali um rosto) e tenta corrigir os pesos de modo a obter resultados mais próximos do desejado. Este processo é repetido inúmeras vezes para cada um dos *inputs* (imagens, por exemplo) que constituem a base de dados de treinamento, de modo que todos "passam" inúmeras vezes pela rede. Na medida em que os *inputs* são repassados à rede, os *outputs* são contrastados com o resultado esperado. A depender do grau de acerto ou erro, o algoritmo tenta refinar os pesos de modo tal que permita à rede reconhecer um *input* sem deixar de ser capaz de reconhecer outro. Ele pode fazer isso, por exemplo, alterando um peso de um dado nodo de 0.8 para 0.89852, o que pode ser interpretado, grosso modo, como a troca de uma regra que diz "neste espaço, é necessário que haja um nariz" para uma regra mais flexível, talvez exprimível como: "em geral, existe um nariz neste espaço, mas há exceções, então é conveniente contrastar com outros resultados antes de concluir".[13] Em certa altura, será possível avaliar o grau de acerto e erro da rede de modo geral. Se o objetivo for fazê-la reconhecer rostos humanos (e não um rosto humano específico), e ela, diante de uma imagem que contenha um rosto humano, reconhecê-lo corretamente em 60% dos casos, será necessário continuar a ajustar os pesos de modo a incrementar esse valor, ao mesmo tempo em que se evitam falsos positivos (casos em que a rede reconhece um rosto onde não existe nenhum). Uma rede confiável precisa apresentar um alto índice de acerto, normalmente acima de 90%.

Evidentemente, um processo deste tipo tem como resultado uma rede pouco útil, exceto talvez para usos muito específicos. Isto porque tal rede se restringirá a reconhecer apenas exemplares iguais àqueles presentes na base de dados utilizada para treiná-la. Contudo, se o objetivo for utilizar redes conexionistas para modelar a cognição humana, é preciso levar em conta que os processos cognitivos típicos são bastante diferentes. O ser humano reconhece qualquer rosto humano como sendo um rosto humano, ainda que nunca o tenha visto. A abordagem conexionista tem condições de abarcar esse fenômeno a partir da capacidade de *generalização*. Esta capacidade está ligada ao fato de que nodos podem trabalhar com valores contínuos. Ainda que, no fim das contas, a rede precise decidir se reconhece ou não um rosto como sendo um rosto (um resultado binário, portanto), o processo decisório acomoda certa

[13] Dada a imprecisão, reitera-se que estas exposições não devem ser tomadas como rigorosamente fiéis ao que se passa, mas sim uma tentativa de fornecer uma noção intuitivamente plausível sem a necessidade de expor detalhes técnicos.

vagueza. Uma rede pode concluir, por exemplo, que um *input* 0.7 está mais para 1 do que para 0, assim como concluir que uma entrada 0.758 está mais para 0.8 do que para 0.7, e assim por diante. Com efeito, torna-se possível lidar com uma margem de variação, permitindo reconhecer um rosto como sendo o rosto de alguém conhecido, ainda que ele tenha trocado de óculos ou deixado a barba crescer, entre outras variações. A capacidade de generalização, portanto, é uma característica essencial para que o conexionismo possa ser utilizado na geração de modelos da atividade cognitiva humana.

Cabe, por fim, esclarecer que o conexionismo não é, por si só, uma teoria da cognição. Ele é melhor compreendido como um *framework* no interior do qual é possível elaborar diferentes teorias. O papel e a natureza das representações mentais admitidas variam, portanto, a depender das especificidades de cada uma. Esse cenário deu origem a uma pluralidade de teorias conexionistas da cognição. Existem, por exemplo, as abordagens eliminativistas, que abrem mão da noção de representação e podem ou não abrir mão da noção de computação. Dentre estas, destaca-se a abordagem baseada em sistemas dinâmicos, tais como a de GELDER (1997), THELEN; SMITH (1992) e CHEMERO (2009). Como o FP é um problema ligado ao uso de representações, abordagens não representacionais não sofrem do FP. Contudo, elas enfrentam seu próprio conjunto de desafios, originando um grande debate acerca do papel das representações em teorias da cognição. Embora este debate esteja fora do escopo desta investigação, ela tem como um dos seus objetivos mostrar que o FP não pode ser utilizado como argumento direto contra o uso de abordagens representacionais, tal como defende, por exemplo, DREYFUS (2007). Este ponto, espera-se, ficará mais evidente no decorrer da discussão.

Existem também abordagens simbólicas, tais como a de HUMMEL (1997) e SHASTRI; AJJANAGADDE (1993). Estas fazem uso do mesmo tipo de representação mental que figura na teoria computacional cognitivista.[14] Neste caso, cada nodo é tomado como representando alguma coisa. Ativar aquele nodo é equivalente a fazer uso de uma representação simbólica na computação clássica. Porém, estas abordagens são acusadas por FODOR; PYLYSHYN (1988) de não agregarem nada de novo à teoria computacional. Na visão destes autores, trata-se de um grande esforço que tem como único efeito mostrar como computações clássicas podem ser implementadas numa abordagem conexionista. Este seria um longo caminho cujo destino é o ponto de partida da GOFAI, portanto. É verdade que, dada a intuitiva proximidade entre redes conexionistas e a estrutura neuronal cerebral, este tipo de trabalho pode viabilizar *insights* acerca de como é possível que o cérebro realize computações. Contudo, ao menos *prima facie*, elas nada trazem de útil à busca por uma solução para o FP.

Abordagens que fazem uso de representações distribuídas são distintas tanto das teorias não representacionais, quanto das teorias simbólicas. Mas o que é, afinal,

[14] Há, inclusive, tentativas aparentemente bem sucedidas de implementar versões neurais de máquinas universais de Turing (GRAVES; WAYNE; DANIHELKA, 2014).

uma representação distribuída? Infelizmente, não há resposta simples. Existe muita confusão e dúvida sobre qual seria a característica essencial de uma representação distribuída. Como GELDER (1992) argumenta, a maior parte do uso que se faz dela apenas supõe que alguma definição seja possível, trabalhando, enquanto isso, com alguma noção incompleta ou preliminar. Mesmo assim, o próprio Gelder realiza um esforço para elaborar uma definição. Para o autor, representações distribuídas possuem como característica fundamental a *superposição semântica*. De modo abstrato, e talvez pouco claro, trata-se da ideia de que um mesmo conjunto de recursos físicos pode constituir distintos estados representacionais ao mesmo tempo. O grau de abstração desta definição pode ser reduzido ao mostrar como isso se realiza no caso de redes conexionistas. Embora haja mais de uma forma de implementar representações distribuídas, há um modo particularmente importante na literatura relacionada, que servirá aqui de exemplo.

Como visto, quando uma rede é treinada para reconhecer um certo conjunto de padrões (um conjunto de faces, por exemplo), o resultado é um conjunto de pesos que lhe permite, diante do *input* adequado, reconhecer tais padrões, isto é, externar o *output* adequado. O que constitui a representação distribuída desses padrões é o *conjunto de pesos* atribuídos aos nodos de uma rede conexionista. A representação não é dada por um símbolo, mas por uma *função* de um *input* para um *output*. Assim, as inúmeras possíveis associações entre o *input* e o *output* de uma rede são representadas pelos mesmos recursos, isto é, pelo mesmo conjunto de conexões entre nodos e seus respectivos pesos.

Com efeito, são as abordagens conexionistas pautadas pelo uso de representações distribuídas que constituem uma alternativa representacional ao sentencialismo. Uma teoria que bem sucedida que faça uso delas seria capaz de mostrar que T4 é verdadeira. Mas por que tal potencial é atribuído às representações distribuídas? A ideia fundamental é que elas não dão espaço para distinções claras entre o que figura de modo explícito e de modo implícito no sistema. Como descreve Haselager, esse gênero representacional constitui uma forma de representação *intrínseca*.[15] Ele expõe esse ponto nos seguintes termos:

> [processos cognitivos] não precisam mais ter a forma de uma reconsideração explícita de todas as estruturas simbólicas e suas interconexões. Alterar o valor de um peso automaticamente influencia todo a capacidade que a rede possui de processar informação. (HASELAGER; RAPPARD, 1998)

Isso significa que o conjunto de pesos de uma rede conexionista pode ser capaz de representar e processar, de modo distribuído, todo o conhecimento que o sistema cognitivo possui. Compare-se isto à abordagens sentenciais típicas, como as vistas

[15] Dizer que a representação é intrínseca não significa dizer que uma representação é idêntica a um dado conjunto de pesos. Um mesmo conjunto de informações pode ser representado por meio de conjuntos de pesos diferentes.

no primeiro capítulo. Nestes casos, diante de quaisquer *inputs*, é necessário derivar explicitamente cada uma das consequências que se encontram implícitas, para que só então seja possível determinar o que é ou não relevante, processo este que depende também da derivação de todas as interconexões com o restante das representações.

No caso das representações distribuídas, contudo, este processo não é necessário. Por não permitir uma distinção clara entre o que está "presente" no processo e o que não está, uma teoria pautada pelo uso de representações distribuídas não permite uma distinção clara entre contexto proximal e contexto distal. Isso porque não há nada explicitamente representado no conjunto de pesos. Ao processar um *input*, todo o conhecimento do sistema cognitivo está "presente" a todo momento, mas sem que se possa dizer que está explicitado tal como seria o caso em sistemas clássicos. Ao mesmo tempo, não há nada implicitamente representado porque nada precisa ser derivado antes de ser utilizado no processo cognitivo. O processamento do *input* é influenciado diretamente pelos pesos. Assim, o uso de representações distribuídas parece poder substituir sistemas com representação sentencial. Uma vez que isto aparentemente evita a emergência do FP, parece razoável aceitar o diagnóstico de que o problema está diretamente ligado ao gênero representacional ou, mais especificamente, ao uso de um gênero que permite uma distinção clara entre o que é explícito e o que é implícito. Se for este o caso, então o requisito da explicitude não é uma característica de teorias computacionais, mas sim de um certo gênero representacional, no caso, o gênero sentencial.

A melhor forma de apresentar, de modo mais preciso, o modo como representações distribuídas podem ajudar na solução do FP é através de um exemplo. Isso permitirá também apresentar, com mais clareza, outras propriedades das abordagens conexionistas que, embora não diretamente ligadas à solução FP, são tomadas como vantajosas em relação à computação clássica. Uma tentativa de elaborar uma teoria cognitiva pautada por representações distribuídas, fortemente inspirada em MCCLELLAND et al. (1987), é encontrada em CHURCHLAND (1989). Em vez de faces ou processos decisórios ultra simplificados, Churchland trata diretamente das circunstâncias em que o sistema cognitivo se encontra. Para maior clareza, pode-se considerar primeiro como isto se daria em criaturas mais simples. Considere-se um *input* de informação visual. Em criaturas mais simples, considerando que ela passe a maior parte da vida em um mesmo ambiente, os estímulos visuais recebidos apresentam menor variação. No modelo de Churchland, esses estímulos são tratados como valores que alimentam a camada de entrada de uma rede conexionista, originando assim um conjunto de *outputs* típicos. Estes *outputs* típicos podem ser localizados num espaço imaginário de *outputs* possíveis, caracterizando assim o que Churchland chama de *hot spot*. Estes *hot spots* descrevem as circunstâncias que o sistema cognitivo tipicamente encontra diante de si.

Churchland supõe que cada contato com um determinado conjunto de circunstâncias pode alterar o registro dos pesos e vieses alocados a cada nodo da rede neu-

ral. Assim, a criatura pode acumular conhecimento do seu ambiente por contato com exemplares das circunstâncias típicas. A cada contato com uma circunstância típica, os pesos podem ser alterados de modo a registrar não apenas a recepção de estímulos visuais, mas também as ações que a criatura realiza ou tenta realizar naquelas circunstâncias. De que modo esta alteração se dá? Churchland não está comprometido com nenhum algoritmo de aprendizado em particular como o *back propagation*, mas supõe que *algum* algoritmo incorporado ao organismo guie o processo, de modo análogo ao que se dá no treinamento de uma rede conexionista artificial. Este ponto, ainda que fundamental, é deixado em aberto. Como resultado, o sistema cognitivo representa, de modo distribuído (isto é, por meio de conjuntos de pesos) informações sobre as circunstâncias, as ações que foram tomadas e os efeitos obtidos em um mesmo *hot spot*, constituindo assim espécie de *protótipo* que pode ser utilizado para que a criatura possa guiar-se em situações semelhantes. Churchland diz:

> A imagem que tento invocar, da vida cognitiva de criaturas simples, atribui a elas uma "biblioteca" organizada de representações internas de várias situações perceptuais prototípicas, situações para as quais *comportamentos* prototípicos são o *output* computado (. . .) (CHURCHLAND, 1989, p. 207)

O modelo conexionista parece permitir explicar como *inputs* refinados podem gerar efeitos complexos nas redes neurais cerebrais, atingindo um estado tal que corresponde ao reconhecimento, por parte do agente, de um conjunto de circunstâncias como sendo uma situação de um certo tipo. Isso é equivalente a tomar um dado conjunto de circunstâncias como a instância de um estereótipo. Churchland demonstra concordar com esse caráter estereotípico do conhecimento armazenado:

> Quero sugerir que estes vetores prototípicos, quanto ativados, constituem o reconhecimento e a compreensão que uma criatura tem de sua situação objetiva, uma compreensão que é refletida no seu comportamento subsequente. (1989, p. 208)

Como as redes conexionistas permitem realizar pequenas generalizações, torna-se compreensível como é possível que o sistema cognitivo reconheça certas situações como familiares, ainda que não sejam idênticas às experiências passadas. Assim como um rosto pode ser conhecido por vários ângulos, um conjunto de circunstâncias típicas também pode. Mesmo quando um *input* não resultar num *output* que se localize no interior destes espaços estereotípicos, é grande a chance que ele se situe nos seus arredores, o que é equivalente a uma situação menos familiar, mas não inteiramente desconhecida. Quanto mais inusitado o *input*, mais longe de um *hot spot*, isto é, mais distante de um estereótipo o *output* se localizará e menos o sistema cognitivo saberá como agir. Isso parece dar conta tanto de situações num sentido vago (situação de ameaça, situação de fuga, demandas de filhotes etc.) quanto de circunstâncias crescentemente refinadas e complexas: situação envolvendo uma ameaça específica,

num cenário específico (desviar-se de um predador num espaço apertado, por exemplo). Este efeito parece harmonizar bem com o comportamento observável de qualquer criatura cognitiva, inclusive o ser humano.

Outro efeito, relacionado à capacidade de generalização e especialmente importante para esta discussão, é algumas vezes chamado de *memória endereçável por conteúdo*. O nome busca destacar um contraste com o modo como a memória é tratada no âmbito da computação clássica. Lá, o acesso a uma representação armazenada na memória depende do conhecimento do "local" em que a informação se encontra. Tal como em uma biblioteca, é preciso saber identificar a sessão e a prateleira em que um livro se encontra para que apenas então se possa ter acesso a ele. Note-se como isto está diretamente relacionado ao aspecto organizacional do FP. Quando o sistema não sabe de antemão o "endereço" de uma determinada informação, ele precisa *vasculhar* a memória, checando os possíveis espaços em que ela pode ser encontrada. A eficiência do modo como esta busca se realiza é efeito direto do modo como a informação é organizada, tal como já discutido.

Como se dá este acesso no caso da abordagem conexionista? Ao usar representações distribuídas para armazenar informações sobre os estereótipos, a rede permite que um dado estereótipo seja ativado a partir de *inputs* parciais (em relação ao que constitui o estereótipo). Um som de folhas se mexendo é um exemplo de *input* parcial no caso de uma criatura que viva em uma floresta. O estímulo é parcial no sentido de que pode ser constitutivo de mais de um estereótipo. No caso do computacionalismo, isto é suficiente para que o FP venha à tona, afinal, a necessidade de encaixar o estímulo em um estereótipo (um contexto proximal) demandaria a capacidade de navegar no contexto distal. Mas como o sistema trabalha por padrões de similaridade, um dado som geralmente estará mais próximo de um ou outro estereótipo. Com efeito, mesmo o *input* parcial pode induzir o sistema cognitivo a ativar um estereótipo de ameaça ou alerta, o que trará à tona também o tipo de comportamento que o sistema exibe nestas situações. O conhecimento adequado está sendo acessado a partir do seu conteúdo, sem que haja necessidade de percorrer outras possibilidades, tal como no computacionalismo clássico. Com efeito, a linha que distingue contexto distal e contexto proximal parece borrada, e o FP, inibido.

Uma terceira propriedade importante dessa abordagem é denominada *degradação gradual*. Num sistema computacional clássico, danos aos mecanismos representacionais podem levar à falhas cognitivas graves ou generalizadas. Contudo, conforme mostram os estudos de casos patológicos, a cognição humana pode apresentar graus refinados de degradação. Pode-se perder a memória de certos aspectos de uma pessoa ou evento, sem que se perca a capacidade de reconhecê-la, por exemplo. Na abordagem sentencial, esta é uma propriedade difícil de explicar, afinal, se uma determinada representação é perdida, perde-se também tudo o que dela deriva. Já no caso da representação distribuída, esta propriedade parece mais facilmente explicável, visto que um dano a um nodo, ou mesmo a uma porção de nodos, não necessariamente

levará à perda massiva de capacidade cognitiva ou memória. Uma distorção em um peso pode gerar distorções cognitivas, mas de modo muito mais refinado do que a perda de uma representação ou axioma em um sistema sentencial.

A partir desse cenário, Churchland defende que o modelo por ele desenvolvido é aplicável também para o ser humano e suas capacidades cognitivas mais elevadas. O autor acredita que uma rede neural suficientemente grande pode representar, no seu conjunto de pesos, todas as situações típicas com as quais ele venha a ter contato ao longo da vida, explicando assim a totalidade do seu comportamento.[16] Com efeito, a teoria conexionista de Churchland aparenta ter vantagens enormes sobre teorias pautadas pelo modelo computacional clássico, especialmente no modo como este lida com o FP. Em virtude do uso de representações distribuídas, ao reconhecer-se como estando em uma situação do tipo X, o agente se reconhece, ao mesmo tempo, como estando em uma situação em que o comportamento adequado é Y. Como todo o conhecimento relevante está representado no estereótipo de modo distribuído, diante de qualquer conjunto de circunstâncias, não há a necessidade de realizar cálculos ou buscas em um conjunto potencialmente infinito de inferências ou de possíveis interrelações entre elas. O resultado preliminar parece ser uma solução ao FP e uma demonstração de que T4 é verdadeira.

2.2.1 Limitações e desafios do uso de representações distribuídas

Como se viu, o uso de redes conexionistas parece promissor e o cenário descrito é bastante animador. Talvez até demais. Tal como nos primórdios da GOFAI, o otimismo excessivo pode fazer com que diversos problemas sérios sejam deixados de lado. Um primeiro risco a se notar é o de que Churchland tenha incorrido em uma versão da chamada *falácia do primeiro passo bem sucedido*. Sua abordagem pode vir a funcionar bem em micro-universos ou casos artificialmente restritos. O reconhecimento facial, por exemplo, ainda que conte com razoável grau de generalização, não deixa de ser um destes cenários. Será isto suficiente para lidar adequadamente com o contexto distal que envolve situações reais e complexas? A resposta depende de o quão real é a possibilidade de utilizar o modelo de Churchland para tratar da cognição humana e o vasto volume de representações que ela parece demandar. Essa questão pode ser desenvolvida a partir da bem conhecida crítica ao conexionismo escrita por FODOR; PYLYSHYN (1988). Nesse texto, os autores argumentam que a cognição humana possui duas características fundamentais: produtividade e sistematicidade.

Dizer que a cognição é produtiva é dizer que ela tem potencial para instanciar uma quantidade potencialmente infinita de estados representacionais. Segue-se disto a possibilidade de instanciar um número potencialmente infinito de diferentes estados intencionais. Não há limites intrínsecos para aquilo que a cognição humana é capaz

[16] Chega-se ao ponto de afirmar que o cérebro é uma grande rede neural, com algo entre 5 e 50 camadas. Evidentemente, isto não passa de uma grande especulação.

de representar. Para Fodor, o único modo de explicar como este potencial infinito pode ser realizado por meios finitos é a partir de um sistema representacional que permita a existência de *composições*. Dotadas desta propriedade, representações mentais elementares podem ser combinadas em elementos mais complexos. Estes, por sua vez, podem ser recombinados em elementos ainda mais complexos, e assim por diante, sem limite determinado.

Por sua vez, dizer que a cognição é sistemática é dizer que a capacidade de instanciar de um certo conteúdo representacional implica a capacidade de instanciar outros conteúdos relacionados. Um exemplo clássico é a crença *João ama Maria*. Se o aparato cognitivo for tal como o do ser humano, a capacidade de ter esta crença implica também a capacidade de instanciar outras crenças, como a de que *Maria ama João* ou *Maria é amada por João*. Tal como no caso da produtividade, pode-se explicar essa característica descrevendo uma relação estruturada e composicional entre as representações de *Maria*, *ama* e *João*. Na crença de que *João ama Maria*, essa estrutura atribui um certo papel para João (aquele que ama) que não é o mesmo papel desempenhado por Maria (aquela que é amada), mas que poderia ser. Possuir a capacidade de atribuir este papel a Maria é possuir a capacidade de atribuir este papel a João. Na mesma linha, WASKAN (2003) tem um exemplo intuitivo de como esta capacidade pode se manifestar no campo da ação: tome-se um cenário com uma porta entreaberta e um balde contendo uma bola, posicionado no topo dela. Um sistema cognitivo capaz de predizer que o balde e a bola cairão caso alguém movimente a porta, deve ser igualmente capaz de usar do mesmo conhecimento para inferir que o balde pode ser usado para carregar a bola através da porta ou em outras situações. Se tais exemplos são acurados, então a sistematicidade, de fato, permeia toda atividade cognitiva humana.

Assumindo esse ponto, Fodor e Pylyshyn lançam a pergunta: podem redes conexionistas que utilizem representações distribuídas tratar adequadamente da sistematicidade? Ambos acreditam que não. De fato, este é um grande desafio para o conexionismo, talvez tão grande quanto o próprio FP. Hummel descreve este desafio nos termos de uma tensão entre flexibilidade e sistematicidade:

> . . . sistemas simbólicos modelam prontamente a sensibilidade a estruturas, mas geralmente falham em demonstrar a flexibilidade exibida pelos seres humanos, enquanto sistemas conexionistas exibem flexibilidade no reconhecimento de padrões e generalização, mas tem grande dificuldade em formar ou manipular representações estruturadas. (HUMMEL, 1997)

A imagem que Hummel tem em mente é uma espécie de gangorra: o conexionismo consegue abarcar a flexibilidade humana, mas tem dificuldade em fazer isso sem abrir mão da sistematicidade. A computação clássica (simbólica), por sua vez, trata adequadamente a sistematicidade da cognição humana, mas ao custo da flexibilidade. O diagnóstico de Hummel harmoniza com o que foi até aqui desenvolvido:

a flexibilidade da cognição humana é dada por sua sensibilidade ao contexto distal, que parece fora do alcance do computacionalismo clássico, uma vez que este está preso ao eliminativismo contextual. Resta desenvolver agora o que caracteriza a dificuldade que o conexionismo tem em tratar da sistematicidade. Pode uma abordagem conexionista evitá-la?

Justificar uma resposta positiva exige demonstrar como uma rede pode aprender o papel estrutural que uma dada porção da representação distribuída realiza, e isso deve ser feito a partir de um conjunto de exemplos daquela representação. Por que a referência a "uma dada porção"? Porque, como visto, representações distribuídas, ao contrário das representações sentenciais, não correlacionam partes do veículo representacional a partes do conteúdo. Uma rede treinada para reconhecer faces humanas, por exemplo, não necessariamente saberia distinguir ou "analisar" um rosto em termos de estruturas (o rosto como "contendo" os olhos, por exemplo). Em síntese, ainda que uma rede reconheça regularidades, ela não necessariamente consegue extrair daí as regras que descrevem sistematicamente essa regularidade. HOLYOAK; HUMMEL (2000) exemplificam essa dificuldade do seguinte modo: suponha-se que um modelo conexionista foi criado (isto é, uma rede foi treinada) para associar o *input* 1 ao *output* 1, o *input* 2 ao *output* 2, 3 ao 3 e assim por diante. Caso a rede receba o valor 4 como *input*, a resposta intuitiva é que o *output* deveria também ser 4. Essa intuição vem da sistematicidade que subjaz os exemplos a partir dos quais a rede foi treinada. Se a rede for capaz de depreender essa sistematicidade dos exemplos, ela deverá ser capaz de responder 4 para o *input* 4, bem como x para qualquer *input* x. Em outras palavras, é preciso que a rede depreenda um conjunto de regras que irão constituir a sistematicidade da atividade em questão, e isto precisa ser feito a partir de conjuntos de instâncias de uso destas regras, presentes na base de dados usada para treinamento da rede. Holyoak desenvolve um pouco mais esse ponto nos seguintes termos:

> [a cognição humana] (. . .) é dependente da capacidade de representar papéis e conectá-los a *fillers*. Esta é precisamente a mesma capacidade que permite composição de símbolos complexos a partir de símbolos mais simples. (2000, p. 10)

O autor refere-se à necessidade de representar uma conexão entre o papel que uma representação desempenham na estrutura e a própria representação, que ele denomina *filler*. Essa conexão entre papéis e representações precisa ser representada de modo independente do papel e do *filler*. Tome-se como exemplo a relação *ama()*, constituída por dois papéis: aquele que ama e aquele que é amado. Tome-se também João e Maria como as representações que irão cumprir estes papéis (isto é, que serão os *fillers*) resultando em uma representação como *ama(João, Maria)*. Em que parte da representação encontra-se a informação de qual papel está sendo cumprido por quem? Isto pode ser convencionado de várias formas. A mais comum delas se dá pela posição que cada elemento ocupa na lista de papéis. Ou seja, na representa-

ção *ama(João, Maria)*, o primeiro item da lista cumpre o papel daquele que ama, e o segundo, daquele que é amado. Dessa forma, a conexão entre o papel e a representação que cumpre esse papel é independente, no sentido de que tanto a relação *ama()* quanto as representações de João e de Maria permanecem formalmente idênticas, ainda que se altere o papel que cada uma delas realiza em qualquer outra instância de *ama()*.

Abordagens conexionistas envolvendo representações distribuídas, contudo, não têm esse tipo de mecanismo à disposição. O desafio, portanto, é duplo: é preciso não apenas mostrar que uma rede é capaz de generalizar regras, mas também que ela é capaz de representá-las de modo independente. MARCUS (1998) acrescenta a isso um terceiro desafio: não basta mostrar que uma rede neural pode ser treinada a fim de apreender e representar um conjunto de regras quaisquer. É preciso garantir que as regras apreendidas sejam aquelas que constituem a cognição humana. Marcus desenvolve esta dificuldade com o seguinte exemplo: considere-se os pares de *input-output*: 2-2, 4-4, 6-6, 8-8. Uma vez que a rede tenha sido treinada a partir destas informações, qual seria o *output* adequado para o *input* 7? Intuitivamente, a regra mais plausível à cognição humana pode ser descrita nos termos de uma função como `f(x) = x`, ou seja, o *output* deve ser idêntico ao *input*. Logo, o *output* adequado é 7. Contudo, para garantir que a rede apreenda tal regra, é preciso um algoritmo de aprendizado que garanta que ela descarte outras possibilidades, tais como: `se x é par, então f(x) = x, se não, f(x) = (x-1)`. No caso desta regra, o *output* gerado pelo *input* 7 seria 6. Dada a base de dados utilizada para o treinamento da rede, estas seriam duas formas diferentes de generalização perfeitamente compatíveis com o conteúdo. Que tipo de critério não arbitrário e não *ad hoc* poderia ser utilizado para fazer com que a rede apreenda uma regra e não outra?

Marcus analisou diversos experimentos, além de ter realizado seu próprio conjunto de experiências empíricas, utilizando variantes do *back propagation* sob uma pluralidade de condições. Os experimentos buscavam comparar o comportamento obtido nas redes com o comportamento de participantes do estudo. Embora tenha conseguido fazer com que a rede generalizasse para além do que figurava na base de treinamento, não foi possível restringir as regras aprendidas àquelas que, presumivelmente, melhor descrevem a cognição humana. O modo como a rede generalizava era consistentemente diferente do modo como os seres humanos que participavam do experimento generalizavam:

> Em um experimento (...), testei uma versão de uma rede recorrente simples (...). Para esta tarefa, apresentei a rede com uma série de sentenças como “uma rosa é uma rosa”, “uma tulipa é uma tulipa” e assim por diante, e então testei o modelo em outra sequência com a mesma forma geral, mas contendo uma palavra nova, “um balde é um ...”. Enquanto seres humanos tendem a completar a sequência com a palavra “balde”, a rede recorrente simples que utilizei não ativava a unidade de *output* correspondente à “balde”. Em algumas replicações, nenhuma palavra foi ativada de modo enfático. Qual palavra era ati-

> vada (se alguma o fosse) dependia fortemente do conjunto de pesos aleatórios com os quais a rede havia sido inicializada. (MARCUS, 1998, p. 262)

De modo significativo, o acréscimo de mais e mais dados ao conjunto de treinamento não parecia surtir qualquer efeito relevante. Este é um ponto importante porque a solução não pode ser dependente de um conjunto específico de dados a partir dos quais a rede é treinada. É preciso mostrar que a sistematicidade nasce, necessariamente, de uma pluralidade de diferentes estímulos. Assim como, presumivelmente, a mesma sistematicidade pode emergir em seres humanos a partir de uma grande pluralidade de *inputs* do ambiente (incluindo aí elementos ambientais e culturais), o algoritmo deve ser capaz de derivar as regras estruturais da sistematicidade, tal como presente na cognição humana, a partir de uma grande variedade de espaços de treinamento. Não seria suficiente encontrar um conjunto que, por coincidência, faça a sistematicidade emergir. Este é um desafio maior do que parece, pois muitas redes geradas para trabalhar em cenários relativamente simples (como as utilizadas em aplicações comerciais ou em experimentos científicos) só conseguiram atingir a acurácia necessária a partir de uma cuidadosa seleção do conteúdo do espaço de treinamento. Isto sugere que a diferença entre o modo como seres humanos aprendem a partir da experiência e o modo como as redes “aprendem” está na computação realizada, isto é, no algoritmo de treinamento, e não no volume de dados ou na amplitude do espaço de treinamento.

Não bastasse, conforme argumenta HADLEY (1994), ainda não está claro sequer se a capacidade de agir em conformidade a certas regras, sejam análogas às da cognição humana ou não, constituem autêntica sistematicidade. É preciso também distinguir os casos em que se obteve mera coincidência estatística pré-arranjada. Para isso, ele sugere que se adote como requisito a capacidade de lidar com a *novidade*, mas de um modo distintamente humano. Ele tem em mente a capacidade de generalizar para situações novas de um modo tal que permita isolar o conhecimento das estruturas sistemáticas como sendo o fator a partir do qual a capacidade se mostrou.

Um exemplo que pode pode ilustrar esse requisito é dado por DREYFUS; DREYFUS (1987). Trata-se de uma situação em que o agente está a decidir em qual jóquei ele vai apostar. É uma situação familiar a ele, que presumivelmente possui um estereótipo incluindo o comportamento adequado e os critérios adequados para uma decisão. De repente, ele obtém a informação de que um jóquei tem alergia a pólen, e de que há flores espalhadas por toda a pista em virtude de haver uma ocasião comemorativa qualquer. É plausível supor que estes fatos são imediatamente reconhecidos como relevantes para a decisão de qual jóquei ele deve apostar. Contudo, o conhecimento necessário para tal, não está nem vagamente ligado ao que constitui o estereótipo da situação em que ele se encontra, ou seja, não é um conhecimento que lhe aparece como relevante em função de familiaridade com experiências anteriores. Ao contrário, o reconhecimento destas crenças como relevantes parece demandar a capacidade de

percorrer as relações sistemáticas possíveis, ainda que não previamente mapeadas por familiaridade. Até o presente momento, nenhuma abordagem conexionista bem conhecida parece capaz de fazer emergir este tipo de capacidade em cenários não previamente circunscritos, isto é, fora de micro-universos. O que se obteve foi, tão somente, generalizações em cenários restritos e fortemente condicionadas ao conjunto de dados que constituiu o espaço de aprendizado. Isto parece confirmar o diagnóstico de Hummel: abordagens conexionistas parecem, de fato, adequadas à identificação de padrões familiares, mas não à estruturação sistemática de elementos.

O que se pode concluir, ainda que preliminarmente, dessa discussão? Por ora, é preciso chamar a atenção para o seguinte ponto: a busca pela sistematicidade, tanto enquanto objetivo mais geral, quanto na forma de objetivos específicos (tal como no problema de garantir que as regras generalizadas sejam análogas às da cognição humana), tende a enfatizar o papel do algoritmo de aprendizado. Isso sugere que a busca pela sistematicidade possa assumir uma forma um pouco mais específica: pode a sistematicidade nascer do acúmulo de exemplos processados nos termos de um dado algoritmo de aprendizado? Com isso em mente, convém voltar a analisar a teoria conexionista da cognição proposta por Churchland. Uma teoria cognitiva precisa explicar (ou, pelo menos, viabilizar uma explicação) de como se dá o processo de aprendizado continuo do agente em sua relação com o mundo, além de processos relacionados, tal como o de revisão de crenças. Se o conhecimento do agente é sedimentado na forma de representações distribuídas, e se estas são constituídas por conjuntos de pesos, então revisar crenças, aprender, é revisar os valores dos pesos. É verdade que, em contraste com a abordagem computacional clássica, alterar pesos afeta todo o conjunto de conhecimento de modo global ou holístico, sem que seja necessário realizar um conjunto infindável de inferências. Contudo, permanece abertas questões cruciais: qual peso deve ser modificado? Para qual valor? Sob quais critérios? De modo sintetizado: como o agente aprende? Churchland, por exemplo, reconhece que não há uma resposta a esta questão quando diz: *"quais fatores governam a mudança neste nível mais fundamental? (. . .) Essa permanece uma questão em aberto"* (1989, p. 243).

Evidentemente, como se trata de uma questão empírica, e a teoria de Churchland é de 1989, poderia ser o caso de que esta questão já pudesse ser melhor tratada hoje. Porém, não apenas ainda não se chegou lá, como há também crescente número de razões para pessimismo. O *back propagation*, algoritmo mais utilizado no decorrer de décadas de pesquisa, é particularmente implausível porque, como discutido, depende de pares pré-determinados de *input-output*, isto é, para cada elemento constituinte da base de aprendizado, é preciso que se conheça de antemão o *output* correto. Alguns autores, como o próprio Churchland, acreditam que algoritmos de aprendizagem não supervisionados, como aqueles inspirados pelas "regras hebbianas" de HEBB (2002 [1949]) podem ser bons guias. Contemporaneamente, contudo, mesmo com o advento do *deep learning* e o desenvolvimento de técnicas avançadas como as que envolvem *redes adversariais generativas*, esta questão permanece em

aberto.[17]

É crucial notar que essa questão em aberto é uma reformulação, em termos conexionistas, daquilo que supõe o computacionalista clássico: o algoritmo correto, somado à estrutura informacional correta, resultará num modelo adequado da inteligência. Nesse sentido, tratam-se de diferentes caminhos para um mesmo fim: um parte da flexibilidade e luta para alcançar a sistematicidade, o outro parte da sistematicidade e luta para alcançar a flexibilidade. Tanto o conexionismo quanto a GOFAI parecem pautar-se por um desafio semelhante diante de si: encontrar o algoritmo que, de algum modo, explicaria a totalidade do comportamento humano. Nesse sentido (um tanto estrito), conexionismo e computacionalismo clássico parecem estar em lados opostos, mas em pé de igualdade.

Diante disso, um último (e grave) problema pode, ao mesmo tempo, reforçar a tese de que as dificuldades relacionadas apontam para o algoritmo de aprendizado e demonstrar mais claramente como a gangorra de Hummel, entre flexibilidade e sistematicidade, se manifesta. Uma teoria conexionista que queira ao mesmo tempo, expressar sistematicidade e evitar o FP precisa explicar como as representações distribuídas podem interagir umas com as outras. No caso das representações sentenciais, não há problema algum a resolver. Como já discutido, elas são facilmente implementadas em elementos simbólicos que podem ser composicionalmente estruturados. Em uma representação exprimível por “a bola esta dentro do balde” é perfeitamente possível identificar qual é o elemento que aponta para o balde, qual aponta para a bola e qual aponta para a relação entre eles. Como isto se daria no caso de representações distribuídas? Evidentemente, não é possível organizar distintos conjuntos de representações distribuídas em distintas redes neurais, nem circunscrever ou escolher porções de nodos em uma rede de modo arbitrário. O resultado seria exprimível numa espécie de diagrama em que cada elemento seria constituído por uma rede neural capaz de representar um dado conjunto de estereótipos, objetos etc. Nesse caso, as relações entre estes possíveis estereótipos seriam constituídas pelas relações entre os elementos do diagrama, e não nas representações distribuídas. O resultado seria uma estrutura de dados tal qual as que se usa no computacionalismo clássico, o que traria

[17] Trata-se de uma rede conexionista com grande poder de generalização para além do espaço de treinamento. Isto é possível porque seu treinamento envolve a geração de conteúdo que é então reinserido no espaço. O sistema aprende a partir de suas próprias generalizações. Elas podem, por exemplo, criar uma face humana inexistente com altíssima resolução, a ponto de confundir quem tente identificá-las como falsas. Como elas conseguem gerar conteúdo, torna-se intuitivamente plausível uma teoria em que, a partir de certos estereótipos inatos simples, este tipo de algoritmo possa gerar estereótipos mais complexos. A abordagem é tão recente (ver RADFORD; METZ; CHINTALA (2015) para um exemplo) que a escassez de materiais bem como o foco em aplicações específicas dos materiais existentes (em detrimento das questões filosoficamente interessantes), não permitem uma análise apurada neste momento. É razoável especular, contudo, que não há, neste tipo de rede, nada que lhe permita, por princípio, superar o problema tal como este foi colocado no texto. Continuam necessários o uso de cláusula *ad hoc* e/ou de uma seleção cuidadosa do conteúdo que constitui o espaço de aprendizado, de modo a induzir o algoritmo para um lado ou outro. Em outras palavras: estas redes realizam o mesmo que as outras já realizavam, mas agora de modo mais rápido e eficiente, facilitando a realização de grandes feitos em domínios específicos (direção automotiva, por exemplo).

consigo a necessidade de resolver o FP: qual seria a organização adequada deste diagrama?

O que parece necessário, portanto, é um modo de identificar, no interior de um conjunto de pesos atrelados a uma rede conexionista, a quais possíveis aspectos da representação distribuída eles se referem. Atender esse tipo de demanda parece ter se tornado mais factível com o advento do *deep learning*, constituído por algoritmos de aprendizado que associam um maior número de camadas ocultas à detecção de estruturas hierárquicas na análise do *input*.[18] Grosso modo, é como se certos subconjuntos de pesos fossem tomados como representando uma parcela estruturalmente relacionada com os outros subconjuntos de pesos (cada subconjunto coincidindo com uma camada oculta da rede conexionista). Assim, um subconjunto (camada) pode ser dedicado à detecção de um nariz, outro subconjunto, do olho esquerdo, e assim por diante, de modo "estruturado".

Contudo, Hummel (1993; 1997) demonstrou que a necessidade de vencer um desafio adicional: quanto mais distribuída a representação, maior é a possibilidade de surgirem erros de conexão.[19] Para compreender essa dificuldade, é preciso relembrar o que foi anteriormente descrito como um dos requisitos fundamentais para implementação da sistematicidade: a rede precisa ser capaz de apreender as regras que permitem estabelecer a conexão entre papéis e *fillers*. Como visto, no caso da relação `ama(João, Maria)`, quando expressa por meio de representações sentenciais, a conexão é aquilo que estabelece o papel que João e Maria terão na relação. No sentencialismo, esta é dada pela ordenação (ou qualquer outra convenção para o mesmo efeito). O primeiro elemento terá o papel de ser aquele que ama. O segundo, o papel de ser aquele que é amado. No entanto, representações distribuídas não dispõem desse recurso, isto é, elas não conseguem representar estas conexões de modo *independente*, tal como no sentencialismo. Uma rede pode ser capaz de representar, ao mesmo tempo, João, Maria, e a relação `ama()`, sem ser capaz de distinguir entre "João ama Maria" e "Maria ama João", visto que essa conexão não está representada independentemente em parte alguma. É preciso, portanto, elaborar alguma outra forma de registrar a conexão entre um papel e a representação que preenche esse papel. Este problema é chamado por Hummel e Holyoak (1993; 1997) de *problema da conexão* (*binding problem*).

Na literatura conexionista, há uma pluralidade de propostas para especificar

[18] Para uma visão panorâmica, ainda que rigorosamente técnica, das abordagens pautadas pelo *deep learning*, ver DENG; YU (2014).

[19] Note-se que o desafio de Hummel é anterior ao advento mais recente do *deep learning* e que não foi localizada nenhuma tentativa direta de verificar se os algoritmos contemporaneamente utilizados seriam capazes de superá-lo. No entanto, é possível sustentar um moderado ceticismo. As abordagens mais recentes têm como características gerais uma maior velocidade e eficiência, permitindo seu uso em novas tarefas, que são muito mais demandantes computacionalmente, originando assim diversos feitos impressionantes. Apesar disso, não demonstram qualquer capacidade fundamentalmente diferente das exibidas pelos algoritmos mais antigos, e o desafio de Hummel não é um desafio pautado pela velocidade ou mesmo pela eficiência do algoritmo. Trata-se de algo que os algoritmos precisam se mostrar capazes de superar, seja de modo eficiente e rápido ou não.

aquilo que, numa rede, pode representar a conexão. Uma abordagem possível é fazer uso de uma representação em conjunção. Assim, em vez de dedicar uma parcela da representação para o João, outra à Maria e uma terceira à relação `ama()`, usa-se uma representação "atômica": *João-ama-Maria*.[20] O problema com esta abordagem é que a estrutura não está sendo representada, mas sim deixada de lado. Não haverá qualquer tipo de relação estrutural entre, por exemplo, *João-ama-Maria*, *Maria-é-amada* ou *João-ama-alguém*. Nos três casos, está-se abrindo mão da informação de que João, Maria e a relação `ama()` referem-se aos mesmos elementos. O resultado pode ser, em diversos casos, um uso coerente, mas não devido à uma representação correta da estrutura, e sim por uma coincidência estatística. Os *outputs* conformes à cognição humana devem então ser garantidos por outros meios, isto é, o próprio algoritmo de treinamento deve, de algum modo, garantir que a ocorrência de uma representação como *João-ama-Maria* se dê concomitantemente à ocorrência de *Maria-é-amada*. Conforme Hummel argumenta, no entanto, há uma tensão empiricamente verificável nesta e em qualquer outra abordagem que faça uso de representações distribuídas: quanto maior a superposição de representações distribuídas, ou seja, quanto mais complexo o conhecimento representado, maior é a chance de que essa estrutura não seja mantida de modo consistente. Estes requisitos estão fundamentalmente em conflito.

De fato, não há abordagem que tenha reconhecidamente superado esse aspecto do uso de representações distribuídas. Esta dificuldade levou a maior parte dos autores que tentaram realizar modelos conexionistas da cognição a aceitar que representações distribuídas não podem ser responsáveis pela totalidade dos processos cognitivos. Com efeito, alguns autores buscaram por alternativas híbridas. Nelas, representações distribuídas e sentenciais partilham espaço.[21] As representações sentenciais são inseridas por meio de uma semântica localista, que atribui um referente a cada nodo de uma rede. Um nodo `a` pode representar Maria, um nodo `b` pode representar João, e assim por diante. Um dos efeitos colaterais desta abordagem, contudo, é que abre-se mão do caráter intrínseco das representações distribuídas. Retoma-se a distinção entre implícito e explícito que caracteriza o sentencialismo: se o nodo está ativo, a representação se faz explícita, caso contrário, a representação se faz ausente ou implícita. Assim, a rendição a abordagens mistas, ainda que promissoras na solução do problema de conexão, trazem consigo todos os problemas do computacionalismo clássico, inclusive o FP. Trata-se, portanto, de um longo caminho, cujo ponto de chegada é o ponto de partida da computação clássica. É verdade que esta linha de pesquisa pode contribuir de várias formas. Ela pode vir a possibilitar *insights* sobre como a computação clássica pode ser implementada em uma rede conexionista tal como a que, supostamente, constitui o sistema neuronal cerebral. Contudo, uma vez que ela nada colabora para uma solução ao FP, todos os caminhos parecem apontar

[20] Este é o caso da abordagem de SMOLENSKY (1990), por exemplo.
[21] Para um exemplo, vide SHASTRI; AJJANAGADDE (1993).

para a gangorra descrita no início desta sessão: as vantagens do uso de representações distribuídas demandam que se abra mão de uma característica presumivelmente fundamental da cognição humana: a sistematicidade.

Dado o exposto, é possível concluir que o caráter intrínseco das representações distribuídas em nada ajuda contra o FP? Talvez, mas isto ainda não pode ser afirmado com segurança. No fim das contas a real natureza da relação entre representações distribuídas e as condições de possibilidade do FP permanece um mistério. Como se viu, há boas razões para pensar que este gênero não é capaz de tratar adequadamente do fenômeno da sistematicidade tal como ela se apresenta na cognição humana. Isso inviabiliza a possibilidade de isolar um cenário cuja única distinção seja a adoção de representações distribuídas, em contraste com um outro cenário em que se tenha adotado representações sentenciais, o que permitiria uma comparação mais adequada. Há, no entanto, uma diferença importante: ao contrário de teorias pautadas pela computação clássica, teorias conexionistas que façam uso de representações distribuídas não constituem um exemplo de eliminativismo contextual. Nelas, a gangorra entre a necessidade de explicitar o conteúdo e a necessidade de tratar meta-dependências contextuais não se coloca, dado o caráter intrínseco das representações. Tudo indica, porém, que essa gangorra é substituída por outra: o preço a pagar pela sensibilidade ao contexto distal é abrir mão da sistematicidade. No caso particular da abordagem de Churchland, o cenário exposto sugere fortemente que ela sofre, de fato, do mesmo problema que sofriam as primeiras abordagens da GOFAI: o sucesso em um micro-universo como fonte de otimismo e a subsequente estagnação diante de uma série de desafios, que se mostram assim que se tenta abranger cenários mais realistas. Isso sugere que a teoria proposta não desvelou, ao menos ainda, qualquer princípio geral da cognição.

2.3 Representações pictoriais

Como visto na sessão anterior, o diagnóstico segundo o qual o FP se origina em virtude de uma característica do sentencialismo (a separação clara entre conteúdo explícito e implícito) deu origem à busca por soluções baseadas no uso de um gênero representacional diferente: representações distribuídas. O cenário para esse tipo de representação, contudo, apresenta seu próprio conjunto de desafios, e não parece haver qualquer indicação clara de que tais desafios possam ser superados. Há, contudo, um terceiro gênero representacional a investigar: são as representações *pictoriais*. Nelas, ao invés de uma estrutura sentencial, há uma estrutura pictorial ou imagética. Fazem parte deste gênero estruturas como grafos, mapas, modelos, imagens, ondas contínuas e outros. Ele tem como característica fundamental o uso de estruturas que são *isomórficas* em relação ao seu conteúdo. Aquilo que representa e aquilo que é representado tem sua relação estabelecida pelo compartilhamento de uma forma abstrata

comum. Um determinado modelo pictorial pode assim representar estados de coisas no mundo (objetos, propriedades e relações entre eles). Note-se que de modo algum isso implica similaridade visual. As representações podem ser visualmente distintas, mas apresentar uma estrutura formal idêntica.

Existe extenso debate acerca do uso de representações pictoriais como veículo do pensamento. A maioria destas abordagens defende alguma versão de um pluralismo, em que certos conteúdos do pensamento são carregados sentencialmente e outros, pictorialmente. Parece haver boas razões para supor que modelos como mapas explicam melhor a capacidade humana de navegação, por exemplo. Não se pretende aqui assumir posição neste debate, mas sim discutir se o FP pode ou não figurar como argumento a favor do uso de representações pictoriais como o veículo exclusivo (ou principal) do pensamento. A questão chave pode ser colocada nos seguintes termos: por que imaginar que o uso de uma estrutura isomórfica será útil na elaboração de técnicas que permitam lidar com o contexto distal e evitar o FP? Apesar da farta literatura disponível sobre representações pictoriais, parece haver poucos autores tratando especificamente desta questão. De modo geral, o principal argumento para considerar esta possibilidade é o mesmo que motivou a discussão acerca de representações distribuídas: representações pictoriais podem constituir um veículo capaz carregar conteúdo sem permitir uma distinção clara entre conteúdo implícito e explícito. Ou seja, é uma nova tentativa de culpar o sentencialismo pelo FP.

Como argumentou HAUGELAND (1987), inspirado por PALMER (1978), a estrutura das representações sentenciais utilizam símbolos para representar relações entre os próprios símbolos. Assim, se há uma relação como `maior_que()` entre Brasil e Portugal, serão necessária três representações: uma para Brasil, outra para Portugal e uma terceira para a relação entre estes países. O mesmo se dá, por exemplo, com as diversas propriedades de um objeto representado (a cor de um martelo demanda uma representação da cor, do martelo, e da relação entre eles). Representações pictoriais permitem fazer com que tais propriedades e relações sejam tratadas não como representações adicionais, mas pelas propriedades e relações existentes entre as próprias representações. Num caso simples, porém ilustrativo, a relação `maior_que(Brasil,Portugal)` pode ser substituída por um veículo em que a representação do Brasil seja fisicamente maior que a representação de Portugal. Convencionalmente, pode-se estabelecer outras propriedades. O esquema representacional pode, por exemplo, assumir que países representados em verde são maiores que países representados em amarelo. Assim, esta informação será carregada pela cor das representações. O isomorfismo é mantido por qualquer um destes meios.

De que modo este tipo de representação pode ajudar a resolver ou evitar o FP? A ideia fundamental é a de que, como não é preciso lidar com representações explícitas de relações e propriedades, quaisquer alterações das representações dos objetos carregarão consigo “automaticamente” as revisões necessárias no restante do modelo. Assim, ao fazer uso de representações pictoriais, não é preciso considerar todos os

possíveis *outputs*: eles serão consequência direta das modificações realizadas sobre as representações. O próprio veículo representacional “constrange” os possíveis conteúdos semânticos que ele pode vir a carregar. Elisabeth Camp descreve este ponto da seguinte forma:

> (representações pictoriais) espelham relações semânticas entre os objetos representados de modo direto: um constituinte sintático próximo ou acima de outro constituinte em uma figura ou mapa espelha a relação de proximidade ou de estar-acima-de entre os objetos representados ou entre propriedades no mundo. (CAMP, 2007, p. 157)

Para uma melhor compreensão de como isto se dá, tome-se como exemplo a imagem abaixo.

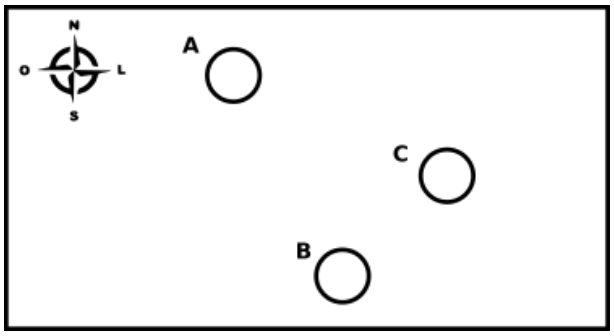

Figura 4 – Exemplo de representação pictorial espacial

A figura expressa uma representação que subjaz uma crença acerca da localização espacial de três cidades: uma cidade A, localizada ao norte, uma cidade B, localizada ao sul e uma cidade C, localizada entre A e B. O contraste com a abordagem sentencial já se mostra: se a crença acerca da localização da cidade A for revista (colocando-a ao sul das demais, por exemplo), não será preciso derivar e/ou calcular explicitamente as relações entre A, B e C. A que se deve tal diferença? Em um primeiro momento, à “automaticidade” com que as consequências de uma revisão na crença se coloca, parece análoga à “automaticidade” com que todas as implicações de um axioma lógico se alteram quanto este é substituído por outro. Como discutido antes, o sentencialismo permite uma distinção clara entre o que é explícito (o que está sendo efetivamente representado no veículo do pensamento) e o que é implícito (tudo aquilo que pode ser derivado, por meio de operações que seguem regras de um sistema formal, a partir do que é explicitamente representado). Já no caso das representações pictoriais, tal como no caso das representações distribuídas, esta distinção não parece possível As relações espaciais da imagem acima, por exemplo, são aparentemente explícitas, nada precisa ser derivado ou calculado para que possam figurar em processos cognitivos. Porém, ao mesmo tempo, uma revisão na posição de um dos elementos não demanda qualquer procedimento adicional, exatamente como seria o caso se um axioma fosse substituído: não é necessário atualizar o conjunto de

derivações possíveis a partir dele. Essa característica, que parece colocar as representações pictoriais em um espaço dúbio entre o que é explícito e o que é implícito, é devida à capacidade de representar relações e propriedades de modo intrínseco.

O defensor do uso de representações pictoriais como parte de uma solução ao FP adere, portanto, tanto à T4, quanto à tese de que representações intrínsecas podem servir como veículo para os processos cognitivos do ser humano. Não faltam exemplos de autores que as tomam como sendo a aposta mais promissora. JANLERT (1996) chegou a defender abertamente que o uso desse tipo de representação poderia fazer parte de uma teoria cognitiva livre do FP, abrindo caminho para uma teoria que explique como o ser humano é capaz de lidar com o contexto distal. Outro autor que defende uma versão dessa abordagem é Johnson-Laird (1980, 2010). Ele faz uso do que chama de *modelos mentais* e defende a tese segundo a qual a razão humana não é fruto da manipulação de representações sentenciais de modo conforme à regras sintáticas, mas sim do uso de modelos isomórficos àquilo que é representado. Outra abordagem, mais recente, é desenvolvida por Waskan (2003, 2006). Ele defende que representações pictoriais podem funcionar como modelos computacionais intrínsecos. Tal como nos demais casos, Waskan acredita que isso se viabiliza porque tais modelos computacionais são isomórficos em relação ao mundo.

Mas afinal, como se daria a explicação do comportamento humano em uma teoria cognitiva que faça uso de representações pictoriais? Para que se mostrem viáveis na explicação de comportamento inteligente, teorias que utilizam representações pictoriais demandam tratamentos distintos dos processos cognitivos. Afinal, se um modelo fosse usado apenas como ponto de partida para realizar inferências, tal como no sentencialismo, então sua adoção não geraria qualquer ganho para o teórico interessado em se livrar do FP. É preciso adotar uma concepção distinta, como a que é sintetizada por Hendricks da seguinte forma:

> A abordagem (...) deveria nos dar uma mão na solução do *frame problem* porque, de acordo com essa abordagem, acessar o conhecimento certo nas circunstâncias certas é tão natural ao agente quanto habitar o mundo e agir como quiser - o modelo mental é paralelo à realidade. (HENDRICKS, 2006, p. 326)

Como tal “paralelismo” seria possível? A ideia geral é a seguinte: um agente que se encontre em uma dada situação possui um modelo mental dela, isto é, uma representação isomórfica desse ambiente. Este modelo pode ser compreendido como um estereótipo da situação concreta. Esse caráter estereotípico constitui um ponto importante, visto que modelos podem ser incompletos, ou mesmo falhos, permitindo assim explicar predições falsas e inferências inválidas. As representações são então utilizadas na realização de “simulações” do mundo, isto é, elas replicam processos mundanos. Ao tentar concluir qual seria o efeito de jogar uma bola contra um quadro, por exemplo, o sistema utiliza, grosso modo, uma representação isomórfica da situação e realiza uma simulação em que a representação da bola é “jogada contra” a

representação do quadro. Como as relações das representações são isomórficas às do mundo real (ainda que de modo falível, dado o seu caráter estereotípico), os efeitos da simulação interna resultam na predição do efeito que a ação terá no mundo. O FP parece inibido, visto que tal efeito se dá sem a necessidade de calcular possíveis variações circunstanciais, e sem que seja preciso circunscrever alternativas relevantes dentre um mar de possibilidades irrelevantes. Em síntese, não há FP.

Se um sistema nestes moldes puder ser desenvolvido, esta parece de fato uma abordagem promissora, com vantagens aparentemente superiores às obtidas pelo uso de representações distribuídas. Parece possível, por exemplo, incrementar o sistema representacional com crescente número de informações, agregando cada vez mais elementos à estrutura pictorial, sem que esta escalada implique a necessidade de recalcular ou revisar relações entre os elementos representados e sem que isto implique dificuldade na representação dessa estrutura. Como sugere Waskan, representações pictoriais conseguem também superar o desafio de explicar a sistematicidade da cognição humana. Para o autor, esta é derivada do próprio mundo:

> . . . basta simplesmente notar que o próprio mundo admite certas variações sistemáticas (e.g. não só o gato pode estar sobre o tapete, mas o tapete pode estar sobre o gato). Em vez de empurrar a estrutura linguística, por assim dizer, “abaixo”, para dentro do veículo do pensamento, proponentes da metáfora do modelo de escala tem toda razão em se queixar que um tratamento igualmente viável de sistematicidade pode ser dado ao empurrar a estrutura do mundo “para cima”.[22] (2003, p. 265)

Em síntese, trata-se da tese segundo a qual, dada uma estrutura representacional isomórfica à uma estrutura mundana, esta apresentará sistematicidade na mesma medida em que a sistematicidade se encontra no mundo. Com isso, Waskan acredita que pode abarcar a sistematicidade do pensamento com a mesma elegância presente na explicação sentencialista.[23] Contudo, se tal estratégia é bem sucedida ou não, é um ponto que não será perseguido aqui. Isso porque, como se verá, a sistematicidade não é o grande desafio das representações pictoriais. Ainda que elas consigam apresentar genuína sistematicidade, elas enfrentam sua própria gama de problemas.

O primeiro desafio é dado por Scott Hendricks. Ele questiona a estratégia de usar representações pictoriais para realizar simulações a partir de exemplos como o seguinte:

> Suponha-se que Jones é um explorador que se encontra face a face com um tigre. Suponha-se que Jones possa rodar uma simulação de Eu-na- atual-circunstância-de-estar-de-frente-com-um-tigre. Ele roda esta simulação usando uma [representação] isomórfica do mundo e um conjunto de mecanismos cognitivos para manipular esta [representação] isomórfica tal como ela se daria no mundo. Por que deveríamos pensar que o resultado de tal simulação será Jones fugindo da cena? Por que o resultado da simulação não poderia ser uma representação de Jones sendo comido pelo tigre? (HENDRICKS, 2006, p. 330)

[22] Waskan chama de “modelo de escala” o que vem sendo aqui chamado de representação pictorial.
[23] A questão da sistematicidade é discutida de modo um pouco mais detalhado por CAMP (2007).

O problema, em síntese, é o de como fazer uso de representações pictoriais em cenários envolvendo razão prática. De que modo tal simulação poderia relacionar-se aos desejos e interesses do agente? Uma primeira possibilidade é postular um mecanismo adicional que interaja com os mecanismos responsáveis pelas simulações. O resultado da simulação seria então avaliado por esse mecanismo, que poderia considerar alternativas a partir do conjunto de desejos do agente. Porém, disto se segue o que o resultado da simulação nada mais é do que uma hipótese que surge no horizonte cognitivo do agente. Nesse caso, não se está diante de uma tentativa de modelar a cognição humana de um modo fundamentalmente dependente do uso de representações pictoriais. Afinal, se o resultado da simulação será avaliado ou interpretado por algum outro tipo de mecanismo cognitivo que irá, este sim, tomar a decisão, então como evitar que este mecanismos sofra do FP? Uma primeira resposta seria insistir que estes mecanismos que manipulam representações pictoriais são, eles mesmos, pautados por representações pictoriais e suas operações são também realizadas por meio de simulações. Contudo, ainda que seja possível implementar algo nestes moldes, isso apenas adiaria o problema. Afinal, continuaria necessário responder ao desafio de Hendricks. Para que esta conclusão possa ser evitada, parece inevitável postular, em algum momento, mecanismos que não fazem uso de simulações. Dado o objetivo de evitar o FP pela adoção de simulações, contudo, isto é equivalente a admitir que houve falha em cumpri-lo.

Waskan realizou uma das poucas tentativas bem conhecidas de efetivamente implementar e rodar simulações desta natureza, com foco em ações práticas (2006). Contudo, não sem surpresa, as simulações se resumiram a pequenos micro-universos, sem qualquer sinal claro de que as técnicas utilizadas podem ser generalizadas para uso em cenários realistas, como aqueles em que é necessário lidar com o contexto distal. Infelizmente, Waskan dá sinais de que não considera este passo sequer necessário. Ele acredita que, ao mostrar simulações bem sucedidas em cenários restritos, o FP já pode ser considerado resolvido, sugerindo que seu nome figura entre aqueles que confundem o FP com seu aspecto inferencial. Com efeito, considerando a escassez de materiais, os problemas conceituais que subjazem abordagens como a de Waskan e problemas de difícil tratamento envolvendo o uso de simulações (tal como o exemplo de Hendricks) sustentam a conclusão preliminar de que esta não é uma abordagem frutífera.

Os problemas do uso de representações pictoriais, contudo, não param aí. SAMUELS (2010) elenca mais alguns desafios que precisam ser superados pelos teóricos. Um deles está ligado à possibilidade implementação. Waskan descreve esse como sendo o problema da migração de uma metáfora explicativa para a descrição de um mecanismo explicativo. No caso das representações sentenciais, seu caráter composicional parece formulado sob medida para facilitar a implementação em termos de mecanismos computacionais. E no caso das representações pictoriais? Waskan admite que ainda não há nenhuma teoria reconhecidamente plausível, e a maior parte

das afirmações a esse respeito são fortemente especulativas. Para o autor, a maioria das abordagens oscilam entre teses incompatíveis com fatos empíricos basilares acerca do cérebro, e teses que não passam de uma reformulação de mecanismos sentenciais, isto é, abordagens que podem ser reduzidas à teorias que utilizem representações sentenciais.[24] Abordagens conexionistas poderiam ser utilizadas também, como sugere Janlert (1996), mas não sem antes superar os desafios que o conexionismo apresenta ao lidar com a sistematicidade, tal como amplamente discutido na sessão anterior. Nesse aspecto, o cenário para representações pictoriais não é nada promissor.

Um segundo problema, também apontado por Samuels, é o de explicar a atividade cognitiva realizada com elementos de domínios não concretos. Conceitos como democracia, propriedades normativas como bondade e justiça, ou ainda propriedades psicológicas como crença e experiência. Qual seria a estrutura destas representações? A que exatamente elas seriam isomórficas? O problema permanece em aberto, com a sensação geral de que os teóricos ainda não conseguiram elaborar um ponto de partida plausível, ao menos não um que tenha se tornado amplamente conhecido.

Cada um destes problemas pode originar amplas discussões, envolvendo extensa bibliografia. Contudo, esta seria uma discussão que, mesmo assim, não traria a resposta realmente importante para a presente investigação. Ainda que todos os problemas acima citados fossem resolvidos, restaria por responder a pergunta: pode afinal o uso de representações intrinsecamente isomórficas evitar o surgimento do FP? Ainda que se suponha que tais mecanismos possam ser descritos com sucesso em uma teoria da cognição, resta aquele que será aqui tratado como o desafio mais diretamente ligado ao FP: como é possível gerar representações isomórficas?

Retome-se a fig. 4 como exemplo. Trata-se de um mapa 2D que permite lidar com certas relações espaciais (e talvez com outras relações que possam ser derivadas destas), mas não há dúvida de que este tipo de estrutura é insuficiente para modelar os processos cognitivos que regem a inteligência humana. É preciso que a representação inclua também algo análogo a relações causais, por exemplo. Suponha-se uma outra interpretação para a fig. 4 em que ela não represente a localização de cidades, mas sim a localização de indivíduos em um dado cômodo. O que aconteceria se A fosse deslocado espacialmente até que se sobreponha a B? A depender da dimensão de origem do deslocamento (no caso de um modelo 3D, A pode vir a cair sobre B), o indivíduo B pode tanto permanecer no lugar, talvez sofrendo algum dano, quanto ter sua posição alterada em virtude da movimentação de A (como no caso de um empurrão).

Considerando um processo cognitivo que faça uso de simulações sobre uma representação isomórfica, quaisquer que sejam as consequências na simulação serão, presumivelmente, as consequências da ação no mundo (ainda que não de modo

[24] O principal argumento contra o uso de representações pictoriais está relacionado a isto, e será melhor desenvolvido logo adiante.

infalível). Com efeito, a representação precisa ser isomórfica a todas as relações e propriedades pertinentes no mundo, pois é isso o que permite que estas sejam intrinsecamente representadas e utilizadas nas simulações. O problema crucial: esse isomorfismo é pressuposto, e não explicado. Quais são as relações e propriedades que devem ser isomorficamente capturadas? Como citado anteriormente, Waskan considera que o uso de representações isomórficas configura solução suficiente para o FP. De fato, o aspecto inferencial do FP é inibido, mas não porque inexiste cálculo a se realizar, e sim porque este já foi feito no momento da criação do modelo, isto é, no momento em que a representação pictorial foi gerada. Ao criar uma representação pictorial artificial, tal como um mapa, seu *designer* especifica quais relações serão mapeadas e de que forma. Contudo, este é precisamente o mesmo tipo de tarefa que é realizada pelo *designer* de estruturas de dados utilizadas em sistemas cognitivos sentenciais. Quais relações devem ser mapeadas? Como lidar com as possíveis variações e nuanças das circunstâncias, de modo que a representação seja, de fato, capaz de ser usada em simulações confiáveis?

A possibilidade de formular estas questões acerca de representações pictoriais sugere que seu caráter intrínseco não lhes é fundamental, mas sim algo derivado de uma estrutura subjacente. Note-se, por exemplo, como é perfeitamente possível descrever todas as relações espaciais possíveis no cenário apresentado na fig. 4 por meio de conjuntos de regras sentencialmente expressas, bastando especificar previamente os movimentos possíveis e as consequências de cada um.[25] Isto é viável em quaisquer cenários suficientemente restritos de modo artificial, mas caso tente-se fazer uso desta estratégia em cenários reais, isto é, casos em que estes conjuntos de regras precisam mostrar-se sensíveis ao contexto distal, o resultado é o aspecto organizacional do FP. O aspecto inferencial não se manifesta, isto é, não há cálculo a realizar no decorrer das simulações, porque todos os cálculos e as possíveis relações e consequências foram previamente delineados de modo a tornar o isomorfismo possível. Ora, a necessidade de mapear todas as relações relevantes, de modo que atendam todas as possíveis nuanças relevantes das circunstâncias, é o que caracteriza o FP.

Nesse sentido, o uso de representações pictoriais não parece constituir uma solução para o FP. Em vez disso, elas permitem apenas uma reformulação do problema. Não por acaso, em todas as simulações realizadas por Waskan, a modelagem das representações foi realizada pelos programadores. Também não por acaso, os testes envolvendo simulações que usem as representações pictoriais se davam apenas em micro-universos artificialmente restritos, objetivando apenas tarefas específicas. Não

[25] Como é típico do meio filosófico, há quem discuta essa possibilidade. Contudo, este é um debate repleto de nuanças envolvendo complicações não essenciais a esta discussão. RESCORLA (2009), por exemplo, defende que há modelos pictoriais não completamente exprimíveis em termos lógicos, mas como o primeiro capítulo exemplificou, o uso da lógica é apenas uma das possibilidades para o sentencialismo. O que se afirma aqui, no entanto, independe de debates sobre quais formalismos são adequados para modelar quais aspectos do mundo. Trata-se apenas de afirmar que o conteúdo expresso por um gênero representacional pode ser exprimido por outro gênero de modo informacionalmente equivalente. Para uma discussão mais detalhada deste ponto, ver Haugeland (1987, 1998c).

é possível garantir, claro, que os programadores jamais virão a encontrar princípios que lhes permitam modelar todas as relações e propriedades relevantes, de um modo flexível e sensível às circunstâncias, permitindo que o sistema lide, inclusive, com o contexto distal em cenários realistas. Mas tal feito seria idêntico ao de resolver o aspecto organizacional do FP. *Ceteris paribus*, diante disto, tanto estruturas sentenciais, quanto pictoriais seriam viáveis. Qualquer que seja a solução para este problema no campo das representações pictoriais, servirá também como solução para abordagens sentencialistas. O FP não serve, portanto, de argumento favorável ao uso de representações pictoriais em detrimento das representações sentenciais. Ambas sofrem do FP, e se for preciso escolher entre elas, esta escolha deve ser pautada por outras razões.

Em síntese, o problema de como modelar relações e propriedades de modo que estas sejam isomórficas é uma reformulação do problema de como estruturar conhecimento sentencialmente representado. O aspecto intrínseco das representações pictoriais não pode ser considerado uma solução para o FP porque esta característica pressupõe uma solução para o aspecto organizacional do FP, isto é, pressupõe o mapeamento prévio correto das possíveis relações entre as entidades mundanas. Dizer que uma estrutura é isomórfica nada mais é do que dizer que alguma estrutura interna ao ser humano tem o arranjo organizacional necessário para evitar que o aspecto inferencial do FP venha à tona. Waskan falha em perceber este ponto, talvez por tomar para si uma compreensão do FP que é limitada ao seu aspecto inferencial. Dado o cenário apresentado, é plausível concluir que representações pictoriais falham em capturar aquilo em função de que o FP emerge, além de não constituírem qualquer avanço por si mesmas.

2.4 Considerações finais

Dada a complexidade de alguns dos temas abordados, cabe aqui uma rápida revisão panorâmica da discussão desenvolvida. Este capítulo explorou os entornos de um diagnóstico bastante comum na literatura ligada à filosofia da ciência cognitiva, aqui representado por T4: o FP é caracterizado por um requisito sentencial. A plausibilidade da tese é devida a uma característica que parece específica a este gênero representacional: a capacidade de permitir uma separação clara entre o que é explícito e o que é implícito, dando origem ao requisito da explicitude. Por sua vez, esse requisito seria o responsável por manter o cognitivista preso ao eliminativismo contextual, espaço que inviabiliza um tratamento adequado do contexto distal. Como se viu, duas alternativas surgiram no horizonte, na forma de dois gêneros representacionais distintos: representações distribuídas e representações pictoriais.

Dentre os que adotam teorias conexionistas da cognição, há uma forte ênfase no uso de representações distribuídas. Estas parecem capazes de mostrar uma sensibilidade adequada ao contexto distal, com espaço para vários fenômenos típicos da

cognição humana: influência das experiências prévias do agente, confiabilidade, degradação gradual, acesso à memória por associação, entre outros. No entanto, essas não constituem resposta fácil ao FP, pois o custo destas características envolve abrir mão da sistematicidade genuína, presumivelmente uma das mais elementares propriedades da cognição humana. Desse cenário, depreende-se uma espécie de gangorra entre a capacidade de apresentar flexibilidade e sensibilidade ao contexto distal, e a capacidade de apresentar sistematicidade. Teorias pautadas por representações sentenciais apresentam sistematicidade sem flexibilidade. Teorias pautadas por representações distribuídas apresentam flexibilidade sem sistematicidade. Até o momento, não há qualquer caminho reconhecidamente seguro para descer da gangorra. Com efeito, T4 permanece em aberto, pois não é possível isolar a origem do FP na capacidade de uma distinção clara entre conteúdo explícito e implícito. Isso inviabiliza um comparativo ponto a ponto entre teorias da cognição pautadas pelo conexionismo e pelo computacionalismo clássico.

No caso das representações pictoriais, os problemas enfrentados parecem ser ainda mais sérios e, talvez, intransponíveis. A tese central que motiva seu uso em processos cognitivos, de que o isomorfismo evita o FP, mostrou-se mais uma instância da falácia do primeiro passo bem sucedido: embora seja possível fazer uso de isomorfismo em cenários artificialmente restritos, isso só é viável nestes micro-universos. Já um isomorfismo aplicável a situações reais precisa ser sensível ao contexto distal, e isso só é possível caso o aspecto organizacional do FP seja resolvido, visto que o isomorfismo pressupõe um mapeamento correto de todas as relações e propriedades mundanas relevantes em cada circunstância. Isso significa também que o caráter intrínseco das representações pictoriais não lhes é fundamental, podendo inclusive ser derivado de conjuntos de regras expressas sentencialmente. Assim, a ideia de que os processos cognitivos podem ser caracterizados por simulações mostrou-se inócua, face ao desafio do FP.

Diante desse cenário, a conclusão preliminar mais razoável é a de que T4 permanece uma questão em aberto. Como nenhuma das alternativas mostrou-se uma solução viável, não foi possível isolar o gênero representacional como responsável pelo surgimento do FP. A natureza do problema permanece misteriosa.

3 Contexto e Racionalidade

No capítulo anterior, discutiu-se a possibilidade de o FP ser consequência direta da adoção de um gênero representacional específico, o sentencial. Porém, tal como a adoção de diferentes formalismos, discutida no primeiro capítulo, a adoção de representações distribuídas ou pictoriais mostrou-se incapaz de solucionar ou elucidar o problema. Isso abre espaço para hipóteses alternativas, segundo as quais o FP tem sua origem em algum ponto ainda mais profundo no interior do cognitivismo. Que ponto seria esse? Fodor talvez tenha sido o primeiro a sugerir que o problema alcança a análise da própria racionalidade. Vários filósofos e cientistas cognitivos, Fodor entre eles (1983), entendem que a régua adequada para medir a inteligência de uma criatura é um ideal normativo de racionalidade. Quanto mais próxima é uma criatura cognitiva desta racionalidade ideal, maior a sua inteligência.

Não é exagero dizer que esse papel atribuído à racionalidade é um dos pilares do cognitivismo, ao menos no modo como este se manifestou na GOFAI. Inclusive, boa parte das discussões do primeiro capítulo podem ser lidas como sendo tentativas de modelar processos cognitivos de modo conforme a um ideal racional. Cabe, portanto, perguntar: não poderia ser o FP fruto da adoção desta racionalidade idealizada como norte? Não poderia o FP ser evitado pela adoção de uma métrica distinta para a inteligência humana? Alguns autores acreditam que sim. Embora haja quem defenda o abandono da racionalidade como característica fundamental dos processos cognitivos, as teses mais discutidas mantêm a racionalidade como guia, mas concebem-na de modo distinto, como no caso da racionalidade limitada (*bounded rationality*).

É um fato trivial que toda criatura cognitiva é finita e imperfeita, isto é, opera sob severas limitações de recursos cognitivos, tais como uma relativamente baixa capacidade de memória e um relativamente reduzido poder computacional. Não é trivial, contudo, especificar o papel que essas limitações devem ter na formulação de critérios normativos de racionalidade, como os que são usados para determinar se uma criatura pode ou não ser considerada um agente racional. Como se viu no primeiro capítulo, Haugeland e Cummins defendem que a atribuição de conteúdo a uma criatura dotada de um sistema cognitivo se dá por meio de uma semântica interpretacional. Pode-se agora acrescentar: e esta interpretação se dá a partir de um ideal de racionalidade. Seja na TRM, seja no caso de alternativas não representacionais como a de Dennett (1989), há um ideal de racionalidade a partir do qual uma criatura pode ser considerada portadora de um sistema cognitivo e constituir assim um agente. Neste capítulo, a tese explorada será a de que diferentes concepções de racionalidade permitem evitar o FP sem abrir mão do caráter computacionalista e representacional da TRM. Antes de discorrer sobre questões específicas, contudo, é preciso afunilar o escopo da discussão para capturar de modo mais preciso a relação entre racionalidade

e FP.

3.1 O que está em jogo na discussão sobre racionalidade

Discutir racionalidade no âmbito da filosofia da mente ou da ciência cognitiva é discutir o que Dietrich e Fields (1995) denominaram *visão racionalista da inteligência humana*. Tal visão caracteriza bem o modo como os pesquisadores da GOFAI compreendiam a inteligência. A TRM, enquanto teoria cognitivista, adota a mesma perspectiva: explicar a inteligência humana é descrever os processos cognitivos por meio dos quais esta é realizada, e estes processos devem ser racionais. Mas o que significa dizer de um dado processo que ele é racional? A resposta depende do modo como a racionalidade é concebida.

Fodor, por exemplo, utiliza o processo de investigação científica como modelo dos processos de fixação de crenças e outros processos não modulares (o tipo de processo que gera o FP). Esses processos têm duas características principais, ambas já discutidas no primeiro capítulo: são isotrópicos e sensíveis a propriedades globais. Eles contrastam com os processos que ocorrem no interior de módulos especializados e informacionalmente encapsulados. Como estes possuem limites fixos ao tipo de informação que conseguem levar em conta, não sofrem do FP. O papel que um ideal normativo de racionalidade exerce nesta leitura explicita-se quando Fodor considera que a distinção entre processos centrais e modulares é intimamente ligada ao grau de racionalidade que pode ser atribuído ao processo:

> . . . processamento cognitivo modular é *ipso facto* irracional. Afinal, por definição, processamento modular significa chegar a conclusões utilizando menos do que toda a evidência que é relevante e que está disponível. (. . .) Apressar-se na lida com os problemas e saltar para conclusões é, então, uma patologia característica de estratégias cognitivas irracionais. (FODOR, 1987, p. 139)

Assim, para Fodor, caracterizar um dado processo como racional envolve medir o grau de acesso deste a todos os fatores que podem vir a ser relevantes no interior do sistema. Ele enfatiza esse ponto ao fazer uma associação direta entre tal caracterização da racionalidade e o FP:

> O *frame problem* vai muito fundo; ele é tão profundo quanto a análise da racionalidade. Deflagrações do *frame problem* são sintomas de processamento racional; se você olhar para um sistema que tem o *frame problem*, você pode assumir que ele é racional, ao menos na medida em que ele não é [informacionalmente] encapsulado. (1987, p. 140–141)

Em síntese: um processo é tão racional quanto sofre do FP. Diante disso, Fodor manifesta seu já discutido ceticismo quanto à possibilidade de se resolver o FP, bem como de qualquer ciência que explique os mecanismos dos processos cognitivos racionais. A conclusão de Fodor pode ser expressa nos termos do ferramental

conceitual aqui utilizado da seguinte forma: um processo é tão racional quanto é sensível ao contexto distal. Como processos computacionais não dão conta de explicar tal sensibilidade, e como não há alternativa explicativa à computacional, segue-se a impossibilidade de uma ciência que explique como são possíveis os processos racionais. É notório, contudo, que a conclusão de Fodor faz sentido apenas num ambiente caracterizado pelo eliminativismo contextual. Isto porque o ceticismo manifesto depende de uma adesão forte ao computacionalismo clássico. O termo "forte" tem seu uso justificado por que Fodor rejeita teorias cognitivas não computacionais, tal como se pôde observar na discussão acerca da abordagem conexionista de Churchland. Essa discussão mostrou, inclusive, que é possível insistir nesse caminho e tentar formular teorias da cognição que não estejam presas à computação clássica, mas que os desafios que elas enfrentam são proporcionalmente grandes.

Esse cenário fez com que alguns autores igualmente indispostos a abandonar o computacionalismo e o representacionalismo (ou seja, dispostos a defender a TRM) tentassem "descolar" a noção de racionalidade, distinguindo-a do modelo fodoriano dos processos cognitivos centrais, podendo assim discuti-la de modo isolado. Dietrich e Fields sintetizaram essa motivação da seguinte maneira:

> O problema com a agenda racionalista é que a racionalidade ideal é incompatível com uma psicologia mecanicista computacional. Consequentemente, qualquer teoria racionalista está comprometida com a visão de que, quanto mais inteligente é o organismo, menos mecanizável ele é, e assim, menos a sua inteligência é passível de explicação. (DIETRICH, 1995, p. 279)

A agenda racionalista que os autores têm em mente é precisamente a agenda de Fodor aqui exposta. Debater a própria concepção de racionalidade permite desdobrar um espaço em que se pode isolar e discutir os efeitos do uso de diferentes noções de racionalidade no âmbito da TRM. Contudo, antes de discutir diferentes perspectivas sobre o papel da racionalidade na descrição da inteligência humana, convém delinear um pouco melhor o escopo em que o tema será tratado. Isso permitirá distinguir de modo mais preciso a presente discussão de outros debates em torno da racionalidade humana.

O termo "racionalidade" pode ser compreendido de pelo menos duas formas, que serão aqui denominadas *axiológica* e *deôntica*. O sentido deôntico de racionalidade tem caráter normativo e geralmente caracteriza a régua por meio da qual uma ação ou agente pode ou não ser considerada racional. Normalmente, o sentido deôntico é também associado ao conjunto de princípios utilizados para predizer o comportamento alheio: ao repassar uma informação qualquer para um interlocutor, há uma expectativa sobre quais seriam as reações adequadas, isto é, as reações compatíveis com o que se esperaria de um agente guiado por princípios racionais. Por sua vez, o sentido axiológico refere-se a um certo conjunto de capacidades que uma determinada criatura precisa apresentar para que possa ser objeto de avaliação no sentido deôn-

tico. Faria pouco sentido exigir um comportamento racional (no sentido deôntico) de uma ameba, visto que ela não detém nenhuma das capacidades racionais (no sentido axiológico) para que seja tomada como candidata à avaliação racional.

No âmbito da ciência cognitiva, o que está em jogo é, evidentemente, o sentido axiológico de racionalidade. Trata-se, afinal, de uma empreitada para descrever não apenas quais são as capacidades do agente que constituem sua racionalidade, mas também para explicar os mecanismos por meio dos quais estas capacidades são realizadas. A questão que emerge desse cenário é: qual a relação entre a concepção deôntica e axiológica? Deve uma constranger a outra? Se sim, em que medida?

Colocada nestes termos abstratos, a questão pode não deixar claro o que está em jogo. Tome-se então, como exemplo, a racionalidade do modo como é compreendida por Fodor. Se aplicada normativamente, então uma criatura só será considerada racional se suas ações forem pautadas por um planejamento otimizado, isto é, se suas ações levarem em conta todo e qualquer elemento relevante a que ela possa ter acesso, seja via memória, seja via percepção. Um processo cognitivo capaz de apresentar este comportamento na prática precisaria de, no mínimo, considerar todas as possibilidades permitidas pelas circunstâncias, a fim de escolher a melhor possível. Ora, como já discutido, o FP é um desafio a qualquer implementação de um tal mecanismo. Dado o eliminativismo contextual que caracteriza a TRM, uma consideração de todas as possibilidades demandaria uma representação da totalidade do contexto distal. Se a racionalidade de Fodor for levada a cabo no sentido deôntico, isto é, se for exigido do agente que seus mecanismos sejam capazes de satisfazer plenamente o que é deônticamente exigido, não haverá uma criatura finita sequer que possa ser considerada racional.

Um exemplo ilustrativo do quão absurda parece esta exigência pode ser encontrado em CHERNIAK (1990). Suponha-se um agente que precisa rever seu conjunto de crenças em função de uma nova informação obtida. Um processo cognitivo que busque satisfazer o critério de racionalidade de Fodor demanda que seja realizada uma checagem por inconsistência. Toda e qualquer inconsistência deve ser encontrada e resolvida, talvez descartando alguma informação anteriormente presente ou mesmo rejeitando a nova informação. Tome-se agora um sistema computacional capaz de realizar cada um dos passos necessários à checagem de consistência na velocidade que um feixe de luz leva para atravessar o diâmetro de um próton. O número de passos necessários, por sua vez, será fruto da quantidade de axiomas presentes no sistema. Segundo os cálculos do autor, checar todas as possíveis inconsistências de um sistema com meros 138 axiomas (um número consideravelmente menor do que seria necessário para qualquer sistema que pretenda simular o comportamento inteligente humano) na velocidade especificada demandaria mais tempo do que a idade estimada do universo.

Contudo, uma concepção diferente de racionalidade poderia amenizar a demanda sobre o tipo de capacidade que um agente precisa possuir de modo a ser

considerado racional e, por consequência, inteligente. Essa é, grosso modo, a tese que caracteriza as teorias da cognição pautadas por uma *racionalidade limitada*. Em oposição a estas, a noção de Fodor será denominada *racionalidade plena*. O ponto central da racionalidade limitada é que os limites dos recursos dos agentes, isto é, possíveis limitações axiológicas, devem ser levados em conta ao caracterizar critérios normativos de racionalidade. Nesse sentido, o exemplo acima ilustra por que os defensores da racionalidade limitada sugerem ser completamente irreal submeter os agentes a um critério de racionalidade que demande a eliminação de inconsistências no processo de revisão de crenças. Nesta estratégia, tanto o agente quanto a métrica de racionalidade a que ele é submetido são idealizados.

É evidente que a ilustração oferecida a partir de Cherniak não constitui um argumento por si só. Seu objetivo é apenas o de exemplificar aquilo que está em jogo no debate com defensores de modelos de racionalidade plena. Afinal, pode ser o caso que, embora o ser humano não alcance a racionalidade plena, ela ainda possa ser adotada como um horizonte idealizado, isto é, como uma métrica através da qual o grau de racionalidade de uma determinada criatura pode ser mensurado. Pode-se também incorporar graus de racionalidade, de modo a acomodar agentes com capacidades menores que a ideal. Quando McCarthy e Hayes buscavam estudar a inteligência "por si mesma" no âmbito da GOFAI, faziam uso de uma estratégia análoga. Eles não estavam preocupados com limitações de recursos cognitivos e, nesse sentido, estavam tratando de agentes ideais.

A ideia de que agentes ideais podem ser utilizados nestes casos, contudo, pode ser colocada em xeque. Uma estratégia comum para questionar o apelo a agentes idealizados envolve o uso do princípio de que *dever implica poder*. Este princípio é frequentemente utilizado no âmbito da ética e da epistemologia para argumentar que não faz sentido exigir do agente algo que ele não é capaz de realizar. No cenário de um acidente veicular em que uma vítima está presa às ferragens, não seria adequado dizer que um indivíduo é moralmente obrigado a remover as ferragens para libertar a vítima se ele não tiver condições físicas para tal. Assim, o princípio pode ser utilizado para estabelecer que limitações factuais devem constranger concepções normativas. No entanto, apesar do apelo intuitivo no campo da moral, não é tão claro se este princípio é válido no caso da epistemologia e no caso da racionalidade. Esta questão, porém, é própria de uma discussão diferente da que vem sendo aqui desenvolvida. Por isso, uma discussão das razões que podem levar à adoção de um ou outro ponto de vista está fora do escopo desta investigação. O objetivo do que se segue não é, portanto, tomar partido no debate, mas apenas apresentar o suficiente para que se possa discutir uma questão específica relacionada: pode a adoção de uma racionalidade limitada em uma teoria cognitiva, contribuir para explicações dos mecanismos que subjazem as capacidades racionais de um agente, a ponto de evitar o FP? Essa é a questão que norteará a próxima sessão.

3.2 Racionalidade limitada como solução para o *Frame Problem*

Dizer que adotar uma concepção limitada de racionalidade como norte ajuda a resolver o FP é equivalente a dizer que o FP é fruto de uma concepção equivocada do que é ser racional, ou talvez ainda, que é fruto de uma concepção errada da relação entre racionalidade e inteligência. É possível discutir, por exemplo, qual a abrangência da racionalidade na explicação da inteligência humana e defender que a razão só é aplicável no interior de um dado contexto proximal, sendo necessário "algo mais" para lidar com o contexto distal. Daniel Andler é um exemplo de autor que defende esse caminho (2000a, 2003), questionando, inclusive, a possibilidade de explicar este "algo mais" em termos mecanicistas. Contudo, este não é o caminho tipicamente adotado por autores que defendem a adoção de uma racionalidade limitada. A forma mais típica pode ser descrita do seguinte modo: a razão guia o agente tanto no contexto proximal quanto no contexto distal, mas ela não é aquilo que se pensa que ela é. Ela é limitada, e não plena. Esse é o caminho que será trilhado no decorrer desta discussão. Ele pode ser sintetizado na seguinte tese:

Tese 5 (T5): Em uma teoria da cognição, a substituição de uma concepção da racionalidade como plena por uma concepção desta como limitada (ainda que não toda ou qualquer versão) pode evitar o surgimento do FP.

Exposta em termos tão genéricos, T5 é uma tese de difícil avaliação. Convém enfatizar, contudo, que não se trata de uma tese segundo a qual toda e qualquer concepção não plena de racionalidade evitará o FP. Trata-se antes de afirmar que o FP está intimamente relacionado à adoção de uma concepção equivocada da racionalidade, e que a adoção da racionalidade limitada adequada pode mitigá-lo, mostrando assim sua real natureza ao mesmo tempo em que o resolve. Basta, portanto, que se formule a concepção adequada. Por isso, ainda que uma discussão profunda de diferentes teorias da racionalidade limitada esteja fora do escopo deste capítulo, convém familiarizar-se com algumas das possibilidades mais discutidas. Duas abordagens serão parcialmente esboçadas: a de John Pollock (2006) e a de Christopher Cherniak (1990). A partir destas, emergirão alguns pontos mais específicos que serão então desenvolvidos no âmbito da *teoria da relevância*[1] de SPERBER; WILSON (1995).[2] Como

[1] Daqui por diante, TR.

[2] Diante da grande quantidade de autores e teorias disponíveis na literatura, a escolha destes carece de alguma justificação. A justificativa para o uso de Pollock é majoritariamente metodológica: por ser a única que não mitiga o papel da representação, sua teoria explicita elementos importantes para a crítica que se construirá na órbita de T5. Estes elementos estão presentes nos demais autores, mas de modo oculto ou mitigado, e a apresentação da teoria de Pollock facilita a construção desse ponto. Por sua vez, Cherniak é um autor que trata diretamente, ainda que sob outro jargão, do problema aqui explorado: a relação tensa entre problemas de relevância e o uso representações. Como se verá, diante de dificuldades análogas às enfrentadas por Pollock, ele conclui que há certos aspectos que não podem ser representacionalmente tratados mesmo na TRM. Contudo, ele não desenvolve nenhuma teoria

se verá, embora a TR não seja tipicamente caracterizada como uma teoria que adota uma racionalidade limitada, ela é considerada, por seus autores, como imune ao FP por negar a noção de racionalidade plena.

Antes de iniciar as devidas apresentações, no entanto, vale a pena retomar rapidamente uma linha argumentativa bastante semelhante e já discutida no primeiro capítulo, onde foram trabalhadas as perspectivas de McDermott, Fodor e Samuels a respeito de T3, aqui replicada:

Tese 3 (T3): Reconhecer que buscamos apenas confiabilidade, e não infalibidade, faz do FP um falso problema para a cognição humana.

Como visto, Samuels defendera que a demanda imposta por Fodor, de que processos cognitivos devem ser sensíveis à propriedades globais (como a relevância), era irreal e deveria ser enfraquecida. Em vez de exigir que uma teoria da cognição explique como o sistema é capaz de exibir essa capacidade, é suficiente exigir que ela descreva apenas uma versão imperfeita e falível dela, de modo a coincidir tanto com a eficiência e acurácia tipicamente exibidas pela cognição humana, quanto com os erros que ela também exibe. Apesar da semelhança, não se trata de um apelo à racionalidade limitada, ao menos não necessariamente. Isto fica perceptível em Samuels quando ele fala sobre o modo como seres humanos realizam abdução, um processo que, presumivelmente, depende de acesso à propriedades globais:

> Não há qualquer razão para assumir que nós sempre computamos com sucesso as propriedades (supostamente) globais das quais depende a abdução. Pelo contrário, é bastante plausível que nós frequentemente falhemos e, além disso, que nós deveríamos ter a expectativa de falha quando demandas cognitivas são grandes demais. Em particular, dado tudo o que sabemos a respeito, pode muito bem ser o caso que a abdução humana falha precisamente sob as mesmas circunstâncias em que a visão [computacional] clássica iria assim predizer. (SAMUELS, 2010, p. 288)

O que Samuels tem em mente é, basicamente, que aquelas circunstâncias consideradas computacionalmente intratáveis coincidiriam com as circunstâncias em que a cognição humana efetivamente falha. É precisamente quando os recursos disponíveis são insuficientes que a mente recorre a estratégias outras: saltos, vieses, heurísticas. Com efeito, Samuels não precisa rejeitar racionalidade plena como guia, mas defende uma postura distinta quanto à capacidade humana de realizá-la. Ele nega o princípio de que dever significa poder, e não está sozinho nisso.

---

alternativa para abarcar estes aspectos, deixando-os em aberto. Quem tenta preencher essa lacuna são Sperber e Wilson, por meio de sua teoria da relevância. Além disso, no caso particular desses autores, há tentativas explícitas de mostrar que o FP não é um problema para sua teoria precisamente porque ela nega a noção de racionalidade plena. Assim, a apresentação e crítica de pontos específicos do trabalho destes autores permite construir paulatinamente o argumento central desse capítulo de uma forma, espera-se, mais rica.

Trata-se de uma perspectiva semelhante à defendida por TVERSKY; KAHNEMAN (1983), segundo quem, ao menos certas porções da atividade cognitiva, são caracterizadas pelo uso de heurísticas. Estas evitam a necessidade de cálculos complexos e, em boa parte dos casos, intermináveis, mas entregam resultados falhos quando contrastados com o padrão normativo da racionalidade plena. A visão defendida por esta linha é a de que o ser humano é apenas parcialmente racional, falhando sistematicamente em fazer uso de processos racionais "plenos". Exposta nestes termos, essa parece uma verdade trivial, mas o que caracteriza essa linha de pensamento não é sua concepção do ser humano como uma criatura parcialmente racional, mas sim que ela retém essa imagem ao mesmo tempo em que mantém uma concepção de racionalidade plena ao fundo. Assim, ainda que esta abordagem sofra do FP (em particular, do seu aspecto organizacional, tal como discutido no primeiro capítulo) não seria válido concluir, a partir dela, que o FP não é um problema relacionado ao ideal constitutivo de racionalidade. Uma tal conclusão precisa envolver a análise de teorias comprometidas com alguma versão de uma racionalidade limitada. Por isso, o debate que orbita T3 não serve de guia para o restante da discussão.

3.2.1 O planejamento local global de Pollock

Um exemplo de teoria que adota uma noção genuína de racionalidade limitada é oferecida por POLLOCK (2006). Sua proposta pode ser compreendida como um tratamento da razão prática feito através de um ferramental conceitual típico das teorias da decisão tal como estudadas no âmbito das ciências econômicas. Nesse jargão, um *plano* é uma resposta para um problema prático qualquer: o que fazer diante de *x*? Como a teoria de Pollock tem caráter computacional, um plano não deixa de ser um conjunto de ações executadas em sequência, isto é, um algoritmo. Para o autor, deixar-se guiar pela racionalidade plena implica adesão a um modelo de decisão caracterizado pelo uso de planos "globais" ou "universais". Neles, virtualmente todos os aspectos da vida do agente precisam ser levados em conta a cada passo, por menor e mais trivial que este seja. Sem isso, nenhum ato, de nenhuma espécie, pode ser considerado racional. Pollock expõe sua insatisfação com esse cenário ao dizer que: *"é computacionalmente absurdo supor que seja preciso planejar o resto da minha vida para que possa decidir o que almoçar."* (2006, p. 169).

O desafio dado pelo aspecto inferencial do FP pode ser expresso em termos muito próximos daqueles utilizados por Pollock para expressar sua insatisfação: como *evitar* que toda decisão, mesmo a mais trivial, deva ser continuamente avaliada à luz de um tal plano universal? Afinal, para evitar que seja este o caso, é necessário circunscrever de antemão o subconjunto dos elementos relevantes à elaboração do plano, naquelas circunstâncias. A TRM, presa ao eliminativismo contextual, precisa representar e considerar todo o conteúdo do contexto distal, e não sabe como fazer isto de modo sensível à sua estrutura. O cenário em que se pintou o problema da sen-

sibilidade ao contexto é, portanto, análogo àquilo que Pollock afirma: a cada decisão, virtualmente todas as possibilidades, em todas as circunstâncias concebíveis devem ser levadas em conta, para que só então seja possível fazer uma escolha racional.

Nota-se portanto, já na largada, uma grande semelhança entre o problema que Pollock deseja resolver e o FP tal como este é aqui compreendido. Sendo assim, o sucesso da empreitada de Pollock pode ser o sucesso na superação do desafio que o FP constitui. Para o autor, o objetivo é encontrar uma forma legítima de abrir mão dessa exigência segundo a qual toda solução para um problema prático deve constituir um plano universal, e é isto que ele busca ao negar a adesão à racionalidade plena. A solução proposta é a substituição desta por um modelo de racionalidade limitada que ele denomina *planejamento localmente global*. Contudo, antes de compreender exatamente o que isso significa, é necessária uma rápida elaboração de alguns pontos basilares de sua teoria.

A rejeição de Pollock à racionalidade plena tem dois pontos fundamentais: primeiro, o comportamento do agente não é guiado pela adoção de planos globais, que seriam considerados planos *otimizados* em função das crenças e desejos do agente. Em vez disso, o agente se guia por planos considerados apenas suficientemente bons. Mas o que seria algo "bom o suficiente"? Qual seria a métrica adequada para determinar tal suficiência? O que Pollock tem em mente é algo inspirado no critério de *satisfação* apresentado por SIMON; NEWELL (1958). Suponha-se que um agente precise decidir a carreira profissional que pretenda perseguir. Se esta escolha fosse guiada por um processo que demanda um planejamento otimizado, o resultado seria irrealizável para qualquer ser humano. Uma escolha racional só seria possível após minuciosa consideração de todos os possíveis cenários, incluindo variantes econômicas, políticas e sociais, bem com as possíveis variações daquilo que seria ou não relevante em cada um dos cenários considerados. Seres humanos não conseguiriam, portanto, coletar e processar o volume de informações necessárias para uma escolha racional. Em contraste, uma escolha satisfatória é aquela realizável a partir do montante informacional que o sistema cognitivo do ser humano tem condições de trabalhar em um dado conjunto de circunstâncias, seja este volume qual for. Neste sentido, os recursos cognitivos do agente (aquilo que constitui sua racionalidade em sentido axiológico) tem um papel a exercer na constituição daquilo que é considerado racional num sentido deôntico. Não se exigirá do agente que ele considere satisfatório planos inalcançáveis em função de limitações factuais de seu aparato cognitivo. Pollock explicita sua adesão a este ponto de vista quando diz: *"tomadores de decisão reais não podem construir ou avaliar planos universais, logo, a teria da racionalidade não pode exigir isso deles"* (2006, p. 170).

O segundo ponto fundamental da alternativa proposta, contudo, apresenta um distanciamento da noção de satisfação de Simon. Em Pollock, esse critério tem natureza *dinâmica*. Ao introduzir o que tem em mente, ele diz:

> Tomadores de decisão precisam trabalhar com o melhor conhecimento à sua disposição e, na medida em que novos conhecimentos se tornam disponíveis, eles podem precisar alterar algumas de suas decisões anteriores. Se um plano melhor for encontrado posteriormente, este deverá suplantar o plano inicialmente adotado. (POLLOCK, 2006, p. 181)

Com efeito, Pollock se distancia de Simon na medida em que o critério de satisfação pode, nos termos do autor, *evoluir*. No caso de Simon, uma escolha satisfatória é definida de um modo relativamente fixo em relação à capacidade cognitiva do agente. Em cada circunstância, haverá um certo limiar que caracteriza uma escolha satisfatória, levando em conta tanto as especificidades da tarefa quanto os recursos cognitivos à disposição. Com efeito, diante de duas alternativas *a* e *b* quaisquer, se ambas atingirem o grau de satisfação adequado à situação, a racionalidade do agente não poderá ser posta em questão caso este prefira *a* em detrimento de *b*, ainda que *b* possa ser uma alternativa "melhor" que *a*. No caso de Pollock, contudo, o caráter dinâmico agrega uma diferença importante: caso um plano melhor se apresente em função de novas informações, o agente racional *deve* escolhê-lo, caso contrário, o preço a se pagar é a desconsideração do agente como um ser racional. Assim, enquanto a noção de satisfação gera a necessidade de estabelecer uma métrica e um limiar para o que pode ser considerado satisfatório em uma dada situação, a dinamicidade deste critério gera uma necessidade adicional: é preciso explicar também de que modo, ou sob quais condições, um dado plano pode ser considerado melhor do que outro.

Prima facie, esta necessidade de um critério adicional não é tão clara. Afinal, supondo a existência de uma métrica qualquer que atribua, diga-se, um valor numérico a um dado plano de ação, bastará comparar o valor que acompanha os diferentes planos e que vença o maior (supondo, claro, que ambos superem um valor mínimo, que seria o critério de satisfação). Contudo, como se verá, a viabilidade da proposta de Pollock reside, em grande medida, na distinção entre estes dois aspectos. Isso porque, enquanto o primeiro permite tratar do comportamento do agente em uma situação específica, é o segundo que permitirá ao agente (assim promete o autor) agir de modo racional sem a necessidade de formular planos globais.

Primeiro, aquilo que os critérios tem em comum: ambos se aplicam sobre o que Pollock denomina *expectativa de utilidade*. Trata-se de uma noção característica de grande grupo de teorias econômicas que tentam prever o comportamento do agente racional. Qual o plano que deve ser posto em prática? O sistema cognitivo do agente atribui a cada plano um valor representativo dessa expectativa, e aquele que possuir o maior deles, é realizado. A fórmula exata em função da qual um dado plano possui uma dada expectativa pode variar enormemente. Ela pode levar em conta, por exemplo, as preferências, os desejos e predições sobre o quão provável é o sucesso da empreitada. Assim, os detalhes podem variar, inclusive, em função de teorias auxiliares que precisam ser adotadas, tais como teorias da probabilidade. Por ora, o importante é enfatizar que a elaboração de uma fórmula adequada para cálculo de expectativa de utilidade

não tem pertinência na presente discussão. Embora seja evidente que uma fórmula inadequada possa colocar a teoria de Pollock abaixo, o que se discute é se ela cumpre o que promete ainda que apresente uma fórmula adequada. Tome-se como exemplo o caso de um plano qualquer sendo avaliado nos termos da racionalidade plena: dada a potencialmente infinita quantidade de casos acerca dos quais a expectativa deverá ser calculada, este é um processo inexequível, independentemente do quão simples seja a fórmula usada. Consequentemente, o que é mais pertinente a esta discussão não é a fórmula em si, mas o papel que ela realiza no interior do sistema cognitivo.

Pollock quer rejeitar a racionalidade plena, isto é, ele quer rejeitar a ideia de que é preciso calcular a expectativa de utilidade de todos os planos concebíveis para que se possa decidir como agir. Para isso, em contraste com tais planos universais, ele introduz o conceito de *plano local* e o de *plano mestre*. O que são, e qual seria o papel de tais planos não-universais? Em qualquer dado momento, um agente possui um conjunto de planos "em execução": escrever até tarde, enviar um e-mail, tomar um remédio, buscar um amigo no aeroporto, e assim por diante. A lista pode também envolver planos mais gerais e com maior número de etapas: vender o carro, alugar um novo apartamento, viajar o mundo etc. Estes planos previamente adotados e parcialmente realizados constituem, num dado momento, uma espécie de pano de fundo a partir do qual as decisões do agente são tomadas. Esses planos individualizados por seus objetivos (enviar um e-mail, por exemplo) são o que Pollock chama de *planos locais*. O amálgama destes planos locais, por sua vez, constitui uma espécie de ferramental por meio do qual o agente lida com o mundo, e é denominado *plano mestre*. Ao sintetizar o que tem em mente, Pollock afirma:

> O plano mestre simplesmente mescla um número de planos locais em um plano único. Cada plano local diz respeito ao que fazer sob certas circunstâncias e, assim, o plano mestre resultante diz respeito ao que fazer sob todas as circunstâncias mencionadas em qualquer dos planos individuais locais. (2006, p. 183)

A ideia de Pollock pode ser mais rapidamente exposta a partir do ferramental conceitual já desenvolvido no decorrer desta investigação: aquilo que o autor chama de plano local é o tipo de processo que se dá no interior de um contexto proximal. Se retomado o jargão da GOFAI, um plano local seria o tipo de planejamento realizado a partir das informações contidas num dado estereótipo de uma situação, não sendo necessário checar nada que não esteja dado naquele estereótipo. No outro extremo, um planejamento universal ou global pode ser descrito como um planejamento que se dá considerando a totalidade do contexto distal. Novamente, se retomado o jargão da GOFAI, um planejamento global é um processo que leva em conta absolutamente todas as informações disponíveis (inclusive por derivação) no interior do sistema computacional. O plano mestre de Pollock se situa, portanto, em uma zona cinzenta entre estes dois extremos, constituindo uma espécie de conjunto de estereótipos representando

contextos proximais. Nesse sentido, ele constitui uma espécie de recorte na totalidade das informações que constituem o sistema.

Essa compreensão do plano mestre como um conjunto de estereótipos já é motivo de alerta, afinal, o eliminativismo contextual é caracterizado precisamente por tomar o contexto distal como um conjunto de contextos proximais, e isso pode rapidamente levar ao FP. A proposta de Pollock, contudo, segue um caminho diferente, ao menos em princípio. Dada a estratégia de abandonar a racionalidade plena, o autor não tenta estabelecer critérios para recortar porções adequadas em um gigantesco contexto distal. Em vez disso, ele tenta mostrar como, a partir do acúmulo e da revisão de planos locais ao longo de sua vida cognitiva, o agente dá forma a um plano mestre. Por isso, Pollock enfatiza a necessidade de tomar o plano mestre como algo cuja forma, ainda que flexível, persiste no tempo:

> Quanto tentando aperfeiçoar seu plano mestre, em vez de jogá-lo fora e recomeçar do zero, o que o agente deve fazer é tentar aperfeiçoá-lo pouco a pouco, deixando a maior parte dele intacta. (2006, p. 184)

Jogar o plano mestre fora e recomeçar do zero é precisamente aquilo que o agente pautado pela racionalidade plena precisa fazer a cada passo. Em contraste, na proposta de Pollock, o que faz de um agente um agente, não é sua capacidade de realizar planos universais, mas a capacidade de formular e manter um plano mestre e guiar seu comportamento no mundo por meio dele. Pollock sintetiza sua estratégia do seguinte modo:

> O único modo pelo qual agentes de recursos limitados podem construir e aprimorar, de modo eficiente, um plano mestre que reflita a complexidade do mundo real é construindo-o ou modificando-o de modo incremental. (2006, p. 184)

Uma metáfora adicional pode ajudar a iluminar o que o autor tem em mente: se a totalidade das informações contidas em um sistema cognitivo fosse uma biblioteca, o plano mestre seria constituído pelo conjunto de livros que o agente tipicamente deixa ao alcance de sua mão na sua mesa de trabalho. Diante da necessidade de resolver um problema prático qualquer, o agente não “limpa” a mesa e inicia seu raciocínio tomando a biblioteca inteira como ponto de partida. Em vez disso, ele se limita àquilo que tem sobre a mesa. Aquilo que ele dispõe sobre a mesa é fruto não de um recorte realizado sob medida para a tarefa atual, mas sim fruto do acúmulo de experiências passadas. Com efeito, o plano mestre pode ser compreendido como uma espécie de seleção prévia entre os contextos proximais mais tipicamente acessados. Ao longo do tempo, novos *inputs* (novas informações, situações etc.), podem fazer com que ele perceba ser necessário buscar livros adicionais na biblioteca, assim como pode fazer com que ele perceba que alguns deles já não são mais necessários, devendo ser retirados da mesa de trabalho. Eventualmente, isto pode ser insuficiente mesmo para gerar um plano satisfatório, mas isso não é um problema para a teoria porque,

como tanto enfatizou McDermott, a cognição humana é falível. Em sua lida cotidiana, contudo, o agente se guia apenas por aquilo de que dispõe à mesa. Nesse cenário, o plano mestre constitui uma espécie de contexto intermediário entre o contexto proximal e o contexto distal. Ele é grande o suficiente para apresentar a flexibilidade com que o ser humano lida com situações típicas em grande variedade de situações, mas pequeno o suficiente para que possa ser devidamente tratado no interior do aparato cognitivo humano.

Pode-se finalmente expor a alternativa que Pollock propõe à racionalidade plena: dizer que uma dada criatura é racional significa dizer que ela é capaz de formular, manter e aperfeiçoar continuamente um plano mestre. Isto é o que o autor denomina *planejamento localmente global*. Ele explicita este ponto da seguinte forma:

> (. . .) uma teoria da escolha racional torna-se uma teoria de como construir planos locais e usá-los para aperfeiçoar sistematicamente o plano mestre. Chamo isto de planejamento localmente global. (POLLOCK, 2006, p. 185)

A plausibilidade do projeto de Pollock reduz-se assim à questão: como se dá a formulação, a manutenção e o aperfeiçoamento de um plano mestre? Em função de quê um livro deve ser mantido, buscado ou retirado da mesa de trabalho? Em quais situações uma alteração ao plano mestre constitui, de fato, um aperfeiçoamento e, consequentemente, um feito racional? Evidentemente, esta capacidade não pode ser descrita nos termos de um plano universal, caso contrário, a teoria retornaria à estaca zero.

A solução de Pollock envolve o uso de cálculos de expectativa de utilidade. O plano mestre, tal como concebido, tem ele mesmo uma expectativa de utilidade. Ele deve ser revisto, portanto, toda vez que uma dada alteração resultar em uma expectativa de utilidade superior à atual. Uma alteração que resulte numa expectativa inferior não pode, portanto, ser considerada um feito racional. Assim, diante de uma nova possibilidade de ação que se mostra (um novo plano local), este pode ou não vir a ser acomodado no interior do plano mestre, a depender da resposta à questão: tal acomodação amplia ou diminui a expectativa de utilidade do plano mestre?

Vale notar que o uso do termo “acomodação” pode ser enganoso, pois pode levar a uma compreensão imprecisa daquilo que realmente ocorre. Acomodar um novo plano local de modo que amplie a expectativa de utilidade inclui, mas não se limita à adição de um “livro sobre a mesa”, isto é, de uma entrada num conjunto de planos locais. Por vezes, essa acomodação só geraria um acréscimo da expectativa de utilidade do plano mestre se feita em conjunto com a adição de planos adicionais ou, talvez, do abandono de outros planos locais. Estes são detalhes que participam da fórmula de cálculo de expectativa de utilidade utilizada.

Assim, na lida cotidiana, este cálculo nunca se dá tendo o contexto distal como pano de fundo, mas sim o plano mestre. Como o tamanho deste é limitado pelos recursos computacionais do sistema cognitivo, o resultado é um processo computacional-

mente tratável. Isso permite adotar, por exemplo, aquilo que é sugerido por Samuels e por McDermott: o cálculo de expectativa de utilidade pode ter caráter heurístico e, nesse sentido, falho. O aspecto organizacional do FP não se manifestará porque tais processos se dão sempre contra um pano de fundo finito e representacionalmente tratável, que é o próprio plano mestre.

Contudo, o que caracteriza o *planejamento localmente global* não é esta lida cotidiana, que pressupõe um plano mestre, mas a capacidade de formular e manter um tal plano. É isto que permite que a lida cotidiana possa ser considerada racional, e é aqui que a plausibilidade do projeto reside. O processo proposto por Pollock é complexo, mas pode ser sintetizado do seguinte modo: primeiro, toma-se o plano mestre atual como ponto de partida. Segundo, toma-se uma possível mudança a este plano mestre. Uma mudança, convém reiterar, não necessariamente se resume ao acréscimo de um plano local, mas pode ser constituída de um conjunto de operações (a remoção de um plano ao mesmo tempo em que se adicionam dois planos locais constitui uma única mudança, por exemplo). Terceiro, computa-se a expectativa de utilidade do plano mestre resultante dessa mudança e, caso o resultado seja positivo, realiza-se a mudança no plano mestre.

A viabilidade desta estratégia depende de uma série de pressupostos, cada um envolvendo questões complexas, inclusive no âmbito da implementação. A questão que interessa, contudo, pode finalmente ser colocada: pode tal teoria evitar o FP? Uma primeira suspeita de que este não é o caso poderia recair sobre o terceiro passo: como é possível computar a expectativa de utilidade do plano mestre sem fazer uso de um processo universal? Contudo, basta notar que este procedimento é do mesmo tipo que o realizado na lida cotidiana: trata-se de um processo que pode ser pautado pelo uso de heurística e cálculos de probabilidade que levem em conta apenas o que está disponível no plano mestre. Assim, um plano mestre bem construído resultará em predições e cálculos razoavelmente acurados, de modo compatível com o desempenho que se poderia esperar de um ser humano.

O verdadeiro desafio está oculto no segundo passo, pois ele envolve a *escolha* de uma possível mudança para consideração. Escolher uma possível mudança envolve a habilidade de formular novos planos locais em função de novos *inputs*. O problema: o conjunto de possibilidades de mudança a considerar não é dado pelo plano mestre, mas pelo contexto distal. Com efeito, há um número potencialmente infinito de possíveis mudanças a considerar. Como selecionar aquelas que devem ser consideradas como candidatas à acomodação junto ao plano mestre antes de calcular a expectativa de utilidade de cada uma delas?

Pollock discute algumas possibilidades que levam em conta os objetivos e expectativas do agente. O problema, claro, é que os próprios objetivos e expectativas estão sujeitos à mudança em função de novos *inputs*. Assim, uma tentativa de resolver o problema apontado no segundo passo por apelo a um plano mestre de ordem superior seria apenas uma versão do regresso infinito de contextos que caracteriza

o modo como todo eliminativista contextual precisa tratar do contexto distal. Pollock reconhece isso como um problema, mas acredita estar diante de uma questão de implementação, e não diante de um problema fundamental. O problema é fatal porque, na ausência de uma resposta adequada, retorna a necessidade de apelo a planos universais. Com ela, retorna o problema de recortar as porções relevantes sem que seja necessário checar todas as possibilidades, uma a uma, ou seja, retorna o FP. A conclusão preliminar, portanto, é de que Pollock não apresentou qualquer resposta que lhe permita descer da gangorra que Fodor descrevera: toda tentativa de solucionar o FP constituirá ou uma solução *ad hoc*, e portanto dependente de um contexto específico, ou uma reformulação do FP, como parece ser o caso.

Não é difícil perceber que este era um resultado bastante previsível no decorrer da apresentação. Desde o início da exposição havia claras dependências não resolvidas e já familiares: como o agente é capaz de formular planos? Como manter um plano mestre adequado sem supor um plano mestre de ordem superior a partir do qual este possa ser formulado? O principal objetivo da apresentação da tese de Pollock, contudo, é estabelecer um norte, que será aqui utilizado como um guia na análise de outras teorias: dentre as teorias da cognição que negam a racionalidade plena, há um aspecto comum, que é a concepção de algum tipo de espaço intermediário entre o contexto proximal e o contexto distal. No caso de Pollock, este espaço é o plano mestre. A abordagem falha porque Pollock falha em mostrar como é possível formular e manter um plano mestre que independa de qualquer tipo de plano universal. Isso não permite, por si só, descartar T5, mas permite um desenvolvimento importante na discussão ao tornar saliente um ponto comum à estratégias que envolvem a negação da racionalidade plena. Se o mesmo tipo de dificuldade se mostrar em outras teorias que façam uso de alguma forma de racionalidade limitada, talvez seja possível justificar uma indução pessimista a respeito desta estratégia. É com isso em mente que se parte agora para uma apresentação rápida da tese de Christopher Cherniak.

3.2.2 A racionalidade mínima de Cherniak

Em sua obra *Minimal Rationality* (1990), Cherniak também advoga contra a racionalidade plena. Para o autor, a racionalidade humana, no sentido deôntico, deve sofrer constrangimentos oriundos de fatos sobre a constituição das capacidades cognitivas humanas. As razões que o levam a defender isso são muito semelhantes às razões pelas quais o FP é tomado como um problema insolúvel.

> A fonte de insatisfação mais importante da ideia de uma condição de racionalidade geral vem de sua negação de uma característica fundamental da existência humana, [a saber], que seres humanos estão sob uma condição de finitude, possuindo limites fixos em suas capacidades cognitivas e no tempo disponível para elas. (1990, p. 8)

Para o autor, uma racionalidade plena parece fazer sentido apenas para um "agente pleno", isto é, um agente que não esteja sob as mesmas limitações que os

agentes humanos. Cherniak não aborda diretamente o FP tal como aqui compreendido, mas seu projeto tem uma ambição aparentemente equivalente: demonstrar como é possível que mecanismos finitos sejam capazes de realizar todas as proezas cognitivas de que agentes humanos são capazes, mesmo quando estas parecem demandar recursos cognitivos infinitos. Assim, o sucesso da empreitada de Cherniak constituiria, no mínimo, um passo importante no estabelecimento de T5 como uma tese verdadeira.

Ele inicia sua argumentação tratando da concepção de racionalidade subjacente à atribuição de atitudes proposicionais a agentes (crenças e desejos). Esta é utilizada rotineiramente por todo ser humano. O comportamento alheio (no caso de agentes racionais) é tanto avaliado quanto predito a partir desta concepção de racionalidade: se uma pessoa possui bens valiosos em um cômodo, e se lhe é dito que está a sair fumaça do local, é razoável supor que ela irá preocupar-se e tomar alguma atitude a respeito, a fim de preservar prioritariamente os bens que lhe são valiosos. Tal predição se dá pela atribuição de desejos (manter seus bens intactos) e crenças (fumaça é indicativo de fogo; este pode destruir bens materiais valiosos; este pode ser eliminado mais facilmente se identificada sua fonte prematuramente etc.).

O que constitui essa concepção de racionalidade a nortear atribuições de crenças e desejos? A estratégia de Cherniak é apresentar duas possibilidades de caracterização destes “critérios de racionalidade”, tomá-las como extremos opostos e então defender uma via média. A primeira delas é a já familiar racionalidade plena. É comum haver casos em que, assumido que um agente possui um determinado objetivo, uma determinada inferência pode ser apropriada ou relevante para a satisfação deste objetivo. Parece plausível supor, nesse caso, que um agente deva fazer uso dessa inferência, de modo que o comportamento resultante possa ser considerado racional. Porém, tal inferência pode estar excessivamente distante. Pode ser necessária uma quantidade imensa de processamento cognitivo a partir do conjunto de crenças à disposição do agente, uma quantidade que pode ser inalcançável, dadas as limitações de recursos inerentes a todo sistema finito. Se a exaustão dos recursos disponíveis impede o agente de derivar a inferência necessária, segue-se que tal sistema cognitivo não pode ser considerado um agente racional, visto que seu comportamento não permite que se lhe atribua qualquer crença ou desejo em conformidade com o ideal adotado. Com efeito, ao usar a racionalidade plena como guia, tem-se um cenário em que um sistema de crenças perfeitamente consistente e capaz de realizar todas as inferências relevantes, é condição necessária para tomar um dado sistema cognitivo como um agente racional. Isso gera uma estranheza enorme, claro, visto que é certo haver inferências fora do alcance da cognição humana, mas também é certo que ninguém deixa de atribuir crenças e desejos a outrem, mesmo que este falhe em realizar inferências relevantes.

No outro extremo, seria possível substituir o ideal de racionalidade plena por um ideal de racionalidade nula, isto é, assumir que nenhuma racionalidade é pressuposta quando atribuindo crenças e desejos a um agente. Seria possível atribuir crenças

somente a partir do que um agente cognitivo afirma publicamente ou assente internamente, por exemplo, sem que se considere necessária a existência de qualquer tipo de relação entre elas, tal como a de consistência. Contudo, esta alternativa torna misteriosa a capacidade humana de predizer comportamento alheio a partir da atribuição de crenças. Se não houvesse qualquer demanda racional sobre as relações entre as crenças, seria impossível realizar qualquer tipo de predição, afinal, o comportamento do agente poderia pautar-se por qualquer crença ou desejo, independentemente das relações desta com as demais crenças que lhe são atribuídas. Não seria possível, por exemplo, confirmar se um agente tem ou não determinada crença a partir do seu comportamento, visto não haver qualquer compromisso entre a posse dessa crença e a ausência de outra. Em particular, importa notar que, fosse pressuposta uma racionalidade nula, a empreitada cognitivista seria inviável: a semântica interpretacional adotada no cognitivismo (discutida no primeiro capítulo), por exemplo, não seria plausível, uma vez que ela atribui conteúdo ao sistema cognitivo a partir das relações entre os *inputs* recebidos os *outputs* que ele produz.

Diante destes dois extremos, a proposta de Cherniak é a de trabalhar com uma noção intermediária, que ele denomina *racionalidade mínima* e sintetiza da seguinte forma:

> Se A tem um conjunto particular de crenças e desejos, A irá realizar algumas, mas não necessariamente todas, as ações aparentemente apropriadas. (CHERNIAK, 1990, p. 9)

Posto nestes termos amplos, a condição é simples: para constituir um agente, A não precisa atender a requisitos tais quais aqueles postos pela racionalidade plena, mas há sim critérios que ele deve satisfazer, ainda que estes sejam menos exigentes. Evidentemente, uma síntese como esta deixa tudo muito vago: o que é aparentemente apropriado? Já que não é necessário realizar todas as ações apropriadas, quantas e quais seriam necessárias? A ideia de Cherniak é que a condição de racionalidade mínima possa servir de guia à formulação de teorias cognitivas, mas para isso, ele precisa apresentar respostas a estas e outras questões. Sua estratégia é razoavelmente complexa e caracteriza-se pelo uso de teorias auxiliares.

Antes de desenvolvê-las, contudo, é importante chamar a atenção para o caráter descritivo da condição de racionalidade mínima. Trata-se antes de um elemento axiológico que de uma tese normativa. Para o autor, é essa condição que estaria por trás das atribuições de de estados intencionais a sistemas cognitivos. Isso não é por acaso. A proposta de Cherniak é a de que os princípios aplicados na predição do comportamento de um agente e os princípios axiológicos da racionalidade são os mesmos. Ele reconhece, claro, a possibilidade de discutir a racionalidade em um sentido deôntico:

> Para nossos propósitos, a tese normativa é esta: a pessoa precisa (somente) realizar todas as inferências adequadas viáveis a partir de suas

> crenças tal que, de acordo com suas crenças, tenda a satisfazer seus desejos. (1990, p. 23)

O que o autor quer negar, contudo, é que mesmo uma tal noção deôntica assim limitada tenha papel a desempenhar em teorias da cognição. Dada a identidade que supõe entre os critérios de atribuição de crenças e desejos, e os critérios axiológicos de racionalidade, Cherniak acredita que a inserção de um elemento normativo, por mais mitigado que seja, traria consigo uma consequência indesejável: a expulsão do ser humano do conjunto de criaturas racionais. Afinal, seres humanos nem sempre realizam inferências ou apresentam comportamento que de fato coadune com seus desejos. Ele erra, se esquece, age sob fadiga e sofre outras interferências que o colocariam excessivas vezes aquém mesmo de um tal critério normativo "mínimo". Sob a luz desse critério, portanto, o ser humano não seria considerado um agente. Com efeito, a proposta de Cherniak é a de que aquilo que deve guiar teorias da cognição são descrições com caráter axiológico. Só assim escapa-se da inexplicabilidade resultante da tentativa de fazer uso da racionalidade plena quando tratando de agentes finitos. Importa mais o modo como uma criatura efetivamente se comporta, do que qualquer concepção normativa de racionalidade, por mais mitigada que seja.

Dito isso, o que gera certa estranheza em relação à estratégia geral de Cherniak é que uma teoria cognitiva não tem por objetivo apenas explicar aquilo em função de que um dado organismo é reconhecido como um agente, isto é, é reconhecido como dotado de crenças, desejos e de um comportamento apropriado em função destas. Uma teoria cognitiva busca explicar também aquilo que um agente tem a capacidade de realizar. Assim, ao abrir mão de uma noção deôntica de racionalidade, identificando condições de racionalidade axiológicas com as condições efetivamente utilizadas quando tentando prever comportamentos, Cherniak pode estar correndo o risco de deixar de fora aspectos importantes da cognição humana.

Como notam Chiappe e Vervaeke, *"o sentido axiológico é crucial, uma vez que ele nos previne de atribuir estados mentais a qualquer coisa"* (1997). Não fossem tais critérios, seria possível, por exemplo, atribuir racionalidade a um sistema aleatório: a depender do número de casos levados em conta durante a análise do comportamento, a formação de uma certa curva probabilística poderia justificar a atribuição de crenças e desejos a um determinado sistema. Cherniak concorda com isso, mas é possível questioná-lo (como de fato o fazem Chiappe e Vervaeke) quando afirma que o único sentido de racionalidade que deve guiar uma teoria da cognição é o axiológico. Afinal, uma concepção deôntica da racionalidade limitada pode ser realizável por agentes finitos. Com efeito, uma teoria da cognição poderia evitar a inexplicabilidade que a adoção da racionalidade plena implica, sem limitar-se a descrever apenas elementos axiológicos.

Essa discussão está ligada a um debate mais amplo sobre se, e como, elementos normativos devem ser considerados em teorias da cognição. O problema com

esse debate é que ele se dá em águas muito distantes, e como se verá, não tem efeitos sobre a questão que norteia esse trabalho. Nos termos expostos, a ideia geral de Cherniak é compatível com o que se tenta checar em T5: pode o abandono da racionalidade plena evitar o FP?

Pode-se então dar sequência ao desenvolvimento das características da condição de racionalidade mínima proposta por Cherniak. Como se verá, o autor substitui o uso de uma norma por um conjunto de teorias psicológicas auxiliares que descrevem o que constitui uma condição mínima de racionalidade. O autor apresenta uma hipótese segundo a qual a própria condição de racionalidade não é uma condição única e geral, mas sim uma composição de condições que envolvem diferentes aspectos da racionalidade.

O primeiro passo de Cherniak é descrever uma *condição de inferência mínima*, constituída por um requerimento heurístico e um requerimento dedutivo. O requerimento heurístico mínimo demanda, de todo agente racional, que ele circunscreva um certo conjunto de inferências apropriadas em relação ao seu conjunto de crenças e desejos. Um sistema cognitivo que atenda tal requisito é um sistema capaz de determinar quais são os caminhos inferenciais que devem ser percorridos. Diante do tocar da campainha, por exemplo, esta é a capacidade de "escolher" se a possibilidade de o som ter sido fruto de uma alucinação será ou não inferencialmente desenvolvida, isto é, se possíveis cursos de ação a partir dela serão levados em conta ou não.

Em paralelo, o requerimento dedutivo mínimo descreve a capacidade de detectar quais caminhos, dentre os possíveis, serão aqueles cujo sistema cognitivo tem mais chances de conseguir desenvolver com suficiência. Assim, ainda no exemplo da campainha, o agente pode perceber que é sim o caso de considerar a possibilidade de estar alucinando (ele pode apostar nesse caminho por ter consigo a crença de que tomou alguma substância alucinógena há pouco tempo, por exemplo). Porém, segue-se a pergunta: o que seria razoável esperar que ele consiga inferir a partir disto? O agente pode carregar consigo uma grande carga de conhecimento relevante ou, melhor dizendo, pode carregar consigo certas porções de conhecimento que, conjuntas, permitiriam a realização de inferências relevantes para decidir o que fazer diante da possibilidade de estar alucinando. Mas até que ponto é razoável exigir que o agente as desenvolva e, de modo tão importante quanto, qual seria o desenvolvimento mínimo esperado? Como se verá, este é o tipo de pergunta que Cherniak espera responder pelo uso de teorias auxiliares.

Ainda sobre a condição mínima de inferência, é importante notar que a condição heurística e a condição dedutiva são equiprimordiais, isto é, não há qualquer prioridade necessária de uma sobre a outra. Não se trata de primeiro decidir o caminho inferencial a seguir e apenas depois decidir até que ponto desenvolver dedutivamente as consequências do caminho escolhido. O que há é uma interação mais complexa, em que ora estão em jogo estratégias heurísticas, ora faz-se uso de estratégias dedutivas. As diferentes estratégias de aplicação de sistemas formais apresentadas no

primeiro capítulo (STRIPS, GPS etc.) constituem exemplos de como essa interação pode ser variada.

O segundo passo de Cherniak é descrever uma *condição de consistência mínima*. Todo agente racional deve apresentar a capacidade de eliminar pelo menos algumas das inconsistências presentes em seu conjunto de crenças. Tal como no caso da condição de inferência mínima, trata-se de um meio termo entre um condição de consistência nula e uma condição ideal. Para o autor, ambas tornariam ininteligíveis as teorias cognitivas que por elas se pautassem. Um requisito de consistência nula impede que se faça qualquer sentido do comportamento de uma criatura. Diante da campainha, o comportamento adotado não permitiria qualquer atribuição de conteúdo, visto não haver qualquer comprometimento entre a posse de certa crença e uma certa ação. Por sua vez, um requisito de consistência idealizado segundo o qual toda e qualquer inconsistência deve ser eliminada, tornaria misteriosa a possibilidade de qualquer agente racional não ideal. Neste caso, seria impossível explicar como o agente consegue decidir o que fazer diante da campainha, sem se perder num processo potencialmente infinito de eliminação de toda inconsistência no conjunto de crenças que guia sua ação.

Tanto a condição de inferência, quanto a condição de consistência mínimas, são exemplos de condições específicas que caracterizam a racionalidade mínima. Como o autor admite, não se trata de uma lista de requisitos exaustiva, mas trata-se de algo que abarca as características mais fundamentais da condição de racionalidade mínima. Ao discorrer sobre como como essa compreensão de racionalidade deve guiar o teórico da cognição, Cherniak defende que estes requisitos específicos devem ser tratados por teorias auxiliares. Não se trata de dizer, contudo, que cada requisito específico terá sua teoria correspondente, mas sim que não é necessário que haja uma única teoria capaz de responder por todos eles. O cenário, portanto, é o de um conjunto de teorias que cobrem diferentes aspectos da racionalidade humana, descrevendo, ao mesmo tempo, as condições de satisfação e o respectivo papel de cada uma delas na explicação da racionalidade mínima.

Cherniak descreve dois tipos distintos de teoria: primeiro, uma *teoria das inferências viáveis* (*feasible inferences*). Segundo, uma *teoria da estrutura da memória humana*. Ele sintetiza o papel da primeira nos seguintes termos:

> (. . .) em situações cotidianas, aquele que atribui [racionalidade] possui uma teoria empírica, um perfil da dificuldade de raciocínios para agentes humanos. Esta teoria provê a informação sobre quais inferências o agente deve realizar, qual sejam, as mais fáceis e mais prováveis. (1990, p. 21)

O ponto chave que uma tal teoria deve conseguir explicar, portanto, é o modo como a viabilidade de certas inferências podem variar em um função de um perfil que, presumivelmente, envolve alguma noção de tipicalidade. Isso pretende dar conta tanto do requisito heurístico quanto do requisito dedutivo. Esse perfil pode ser constituído

tanto por predisposições inatas quanto por habilidades aprendidas. Ele é utilizado para atribuir uma espécie de valor preditivo a cada inferência, indicando, em cada circunstância, aquelas que seriam tomadas como apropriadas, e aquelas que não seriam exigíveis. Ao mesmo tempo, este perfil parece incluir uma noção vaga do custo cognitivo de cada tarefa: caso um agente receba a informação de que foram colocadas duas maçãs em um cesto e, posteriormente, receba a informação de que nele já havia três outras maçãs, a inferência de que agora há cinco maçãs no cesto mostra-se viável, gerando então a expectativa de que qualquer agente racional conseguisse fazê-la. A mesma expectativa não se dá, contudo, se a inferência envolvesse centenas de maçãs em centenas de cestos.

Convém enfatizar que uma inferência viável é uma inferência que deve ser realizada em uma determinada circunstância, mas não necessariamente em toda e qualquer circunstância: *"a dificuldade de uma inferência depende não apenas da inferência em si, mas também das condições sob as quais ela é realizada."* (1990, p. 30). Uma criatura é racional, portanto, quando realiza as inferências adequadas em cada circunstância, e cabe a uma teoria das inferências viáveis explicar os mecanismos por meio dos quais isto se dá.

Em Cherniak, o propósito maior de uma teoria das inferências viáveis é fornecer as bases para que teorias cognitivas descrevam os mecanismos a partir dos quais estes perfis se formam e são aplicados. O modo como isso se viabiliza é profundamente dependente, contudo, de um segundo tipo de teoria auxiliar que trata da estrutura da memória humana. Nela, são distintos dois mecanismos elementares: o primeiro focado na memória de trabalho de curto prazo, e o segundo na memória de longo prazo, e estes possuem modos de operação distintos. Esta distinção possui consequências mais profundas do que pode parecer *prima facie*. Diante de uma criatura supostamente racional, seu comportamento é avaliado a partir de um subconjunto de crenças que lhe são atribuídas. Estas crenças constituem sua memória de trabalho, isto é, de curto prazo, e é a partir deste subconjunto que suas ações podem ser avaliadas. Nesse sentido, a tentativa de predizer o comportamento de uma criatura racional inclui uma tentativa de identificar o conteúdo de sua memória de curto prazo. Assim, aquilo que é tomado como racional ou como irracional se dá não pela manutenção de um corpo global e coeso de crenças (caso da racionalidade plena), mas sim em função daquilo que se considera um conteúdo adequado ou não, para a memória de curto prazo. Cherniak diz:

> Quando prevendo o comportamento de uma pessoa, é muito útil ter uma ideia de quais crenças serão processadas e quando. Em particular, uma inferência útil é mensurada em termos de quando as crenças que são suas premissas e regras são simultaneamente "ativadas", ou consideradas em um dado momento pelo agente. (1990, p. 21)

É importante notar como permanece fundamental a dependência das circunstâncias. O conteúdo adequado para uma memória de curto prazo varia em função

do que é típico e do que é particular a cada circunstância Esse tipo de raciocínio é muito semelhante ao utilizado no âmbito da IA quando se fala em estereótipos e, de fato, um dos pontos essenciais da teoria da memória humana de Cherniak é o que ele chama de *compartimentalização*, cujo significado é o mesmo. Isso demonstra o quão diferentes podem ser os aspectos particulares exigidos por uma condição de racionalidade mínima que envolva modos de acesso à memória de curto e longo prazos. Este é também um exemplo evidente do tipo de constrangimento que a condição mínima de racionalidade, tomada de forma geral, exerce sobre teorias cognitivas: uma teoria cognitiva que não trate adequadamente da distinção entre memória de curto e longo prazos não será uma teoria adequada da cognição humana.

Convém enfatizar que Cherniak não está afirmando que estas duas teorias esgotam tudo o que há para se dizer acerca daquilo que constitui uma condição de racionalidade mínima. As teorias apresentadas são *"... exemplos salientes da ampla gama de teorias psicológicas cognitivas de fundo nas quais as condições de racionalidade mínima se embutem"* (1990, p. 22). Com efeito, a essência de sua tese é sugerir a troca de um ideal pleno e simples, mas inalcançável, por um conjunto de requisitos cognitivos mitigados e constrangidos por teorias psicológicas. Tais teorias devem ser, para Cherniak, inspiradas em descrições daquilo que guia as pessoas quando atribuindo crenças e desejos a outrem, buscando fazer sentido de seu comportamento enquanto agente racional.

Apesar de enfatizar essa pluralidade e a possível complexidade da condição de racionalidade, é possível identificar, no projeto de Cherniak, um ponto central a partir do qual suas teses orbitam: ser racional não é mostrar-se capaz de perseguir um norte estipulado nos termos de uma racionalidade plena, mas sim mostrar-se sensível àquilo que é *relevante* em cada circunstância. O próprio Cherniak sintetiza suas ideias de modo semelhante quando, acerca da teoria das inferências viáveis, ele afirma:

> ... somente um pequeno subconjunto das inferências adequadas que seriam possíveis para um agente fazer na prática seriam positivamente úteis para ele em um dado momento. Uma inferência pode ser adequada, mas pode não ser razoável, visto não ter valor preditível no momento e impedir o agente de usar seus limitados recursos cognitivos para fazer outras coisas que são obviamente valiosas no momento. (1990, p. 24)

O mesmo ponto pode ser visto de modo igualmente claro quando este trata da relação entre as memórias de curto e longo prazos.

> [Uma] pessoa não pode agir de modo minimamente racional (...) exceto se, ao menos algumas vezes, as crenças "corretas" sejam recordadas na memória de curto prazo. As crenças corretas aqui são aquelas relevantes na tomada de uma decisão sobre levar adiante ou não uma dada ação. (1990, p. 61)

Assim, não é necessário apresentar uma análise detalhada destas teorias para perceber que a relevância emerge como um ponto central a partir do qual elas se

desenvolvem. Em toda e qualquer circunstância, há um subconjunto relativamente pequeno de inferências que são realizáveis e, ao mesmo tempo, relevantes ao agente. Este subconjunto está longe de ser caracterizado como "perfeito" dadas as circunstâncias, isto é, não se trata de garantir a relevância. Em vez disso, ele é tomado como "bom o suficiente", exatamente como parece ser o caso quando se busca determinar se uma criatura é racional ou não. A relevância aparece em Cherniak como a chave da racionalidade. Em particular, dada a co-dependência entre as duas teorias, a proposta de Cherniak tem o mérito de tornar explícita uma conexão clara entre o modo como um sistema cognitivo realiza inferências (teoria das inferências viáveis), o modo como a informação é organizada (teoria da organização da memória) e a racionalidade. Com efeito, há uma visível analogia entre uma teoria das inferências viáveis e o aspecto inferencial do FP, assim como há uma analogia entre uma teoria da estrutura da memória humana e o aspecto organizacional do FP. A proposta de Cherniak estabelece uma íntima relação entre o modo como a informação está organizada na memória de longo prazo e a capacidade cognitiva de "pinçar" nela precisamente aquela porção do conhecimento que é relevante às circunstâncias. Nesse sentido, a tese proposta é uma síntese do que se vem discutindo desde o começo desta investigação: o modo como o sistema cognitivo humano opera é dependente não apenas de um certo conjunto informacional, mas também do modo como este conjunto de informações se estrutura. O papel que Cherniak atribui à memória de curto prazo é muito semelhante ao papel que Pollock atribui ao plano mestre. A memória de curto prazo pode ser então compreendida como um conjunto de contextos proximais ali presentes em função de sua tipicalidade. Cherniak reconhece a relação entre o modo como a memória de curto prazo se estrutura e aquilo que é típico à criatura cognitiva quando diz que *"... em alguma medida, o modo como os itens devem ser organizados depende das questões que a criatura é mais propensa a fazer, dadas as suas crenças e objetivos."* (1990, p. 66). Uma vez que a informação adequada esteja devidamente mapeada nos mecanismos que constituem a memória de curto prazo, quaisquer processos computacionais seriam exequíveis como se fossem informacionalmente encapsulados. A memória de curto prazo deve ser então preenchida de modo compartimentalizado, isto é, por estruturas estereotípicas que constituem perfis de uso das informações. Pressuposto isso, basta que este espaço seja percorrido por processos capazes de honrar estas estruturas estereotípicas. O modo exato de como a organização e o exercício de tais perfis inferenciais se dá, é objeto tanto de uma teoria das inferências realizáveis quanto de uma teoria da organização da memória, ou seja, é papel destas teorias indicar de que modo se constitui o contexto proximal em que a atividade cognitiva é realizada.

A tese de que a racionalidade deve ser explicada não em termos de um norte único a perseguir, mas de um conjunto de habilidades é, de fato, poderosa. Contudo, é possível argumentar que as teorias desenvolvidas por Cherniak permanecem carentes de uma explicação adequada para a sensibilidade ao contexto. A questão ainda não foi tratada em seus aspectos mais fundamentais, tal como a determinação da relevân-

cia. Cherniak poderia talvez responder que o tratamento adequado da relevância deve ser objeto de uma terceira teoria auxiliar, a ser utilizada em conjunção com a teoria das inferências viáveis e com a teoria da estrutura da memória. A esse respeito, Cherniak parece até mesmo perceber os problemas ligados ao que fora aqui denominado eliminativismo contextual:

> . . . inferências aparentemente apropriadas não podem ser selecionadas somente por meio de inferênciais práticas reais - conscientes ou não - ou haverá um regresso. Como um primeiro passo para evitar este problema, podemos dizer que em muitos casos o agente não realmente decide levar a inferência adiante. Em vez disso, a ação de inferir precisa apresentar ampla conformidade a desejos e crenças por meio de mecanismos fixos de seleção ou orientação não integrados, não conscientes, e que não envolvam processos racionais de qualquer tipo. Tais mecanismos podem ser adquiridos, por exemplo, pelo aprendizado de "estilos cognitivos" - ou a seleção natural pode ter "projetado" o agente de modo tal que, sendo um organismo eficiente, ele leve a cabo certas inferências. (1990, p. 12)

Grosso modo, o que Cherniak sugere é a existência de mecanismos que participam do aparato cognitivo não restritos àquilo que seria admissível no âmbito da TRM, dado o compromisso desta com o eliminativismo contextual. Em outro momento, ele chega a especular sobre a existência de *"pré-processadores heurísticos"* (1990, p. 118) que seriam inatos e responsáveis por evitar que o organismo se perca em processos inferenciais tão irrelevantes quanto infindáveis. No entanto, a ideia de que mecanismos cognitivos podem ser constrangidos por mecanismos inatos não é de todo estranha à TRM. Os próprios processos algorítmicos são, num certo sentido, constrangimentos ao que um organismo pode ou não realizar com a informação, podendo tanto evitar que certos caminhos inferenciais sejam trilhados quanto carregar conteúdo implícito, tal como discutido no capítulo anterior. Mecanismos inatos deste tipo podem ajudar a explicar por que, por exemplo, um ruído súbito aparece ao organismo como mais demandante de atenção do que um ruído contínuo, mas haverá um longo e misterioso caminho a percorrer se o objetivo for explicar a flexibilidade da inteligência humana face ao contexto distal. Muito provavelmente, a ideia de "estilos cognitivos" aprendidos poderia ser utilizada para dar conta dos contextos a que o ser humano se mostra sensível de modo não inato. O problema é que esta é uma ideia profundamente obscura, e Cherniak não dá qualquer passo mais substancial para explicar seus mecanismos. Ao contrário, ele por vezes parece apenas pressupor que ela não seja problemática. Ao falar sobre como a estruturação da memória pode se dar, ele diz:

> Para seres humanos, algumas das características básicas desta estruturação devem ser resultado da seleção natural, uma vez que teriam sido úteis à sobrevivência em qualquer ambiente tal qual o terrestre (. . .). O resto do esquema organizacional particular do indivíduo é aprendido, uma parcela como parte da cultura, mas muito na forma de hábitos cognitivos idiossincráticos e flexíveis, baseados na experiência passada. (1990, p. 66)

Este permanece, portanto, como um ponto subdesenvolvido em sua teoria e, talvez, até mesmo um ponto cego, se for pressuposto que a viabilidade de uma teoria cognitiva capaz de explicar o comportamento inteligente está entre os objetivos do autor. Não é difícil demonstrar, de modo mais preciso, um lugar onde esse ponto cego se apresenta nos mecanismos que permitem o trânsito entre a memória de longo e de curto prazos. A ideia do autor é que, dada uma crença *r*, se esta for implicada por *p* e *q*, então ela só deverá estar presente no interior da memória de curto prazo se, primeiro, as circunstâncias forem tais que *p* e *q* figurem como relevantes naquelas circunstâncias e, segundo, se ela for plausível, visto que *r* pode ter uma relação bastante remota (em termos de passos inferenciais necessários) e pode ser vista como tendo um custo cognitivo alto demais para que seja razoável exigi-la. Isso equivale a dizer que o ser humano é, ao mesmo tempo, capaz de manter um conjunto de recursos informacionais minimamente consistentes, mas trabalhar com eles de modo não limitado à tais relações de consistência (visto que disto se originaria o aspecto inferencial do FP). Ora, num cenário pautado pelo eliminativismo contextual, isto faz com que emerja de imediato a questão da organização da memória de longo prazo: *r* precisa estar facilmente acessível para ser transposta para a memória de curto prazo em alguns casos, mas não todos. Este é precisamente o mesmo problema encontrado no fim do primeiro capítulo: na TRM, perguntar pela estrutura da memória de longo prazo é perguntar pela estrutura do contexto distal.

Além disso, como se viu no capítulo anterior, para que uma condição de inferência e consistência mínimas possam ser satisfeitas, não basta que a organização seja constituída por um acúmulo de perfis ou compartimentos gerados a partir da experiência do agente ou de qualquer outro modo arbitrário que não atenda aos requisitos mínimos de inferência e consistência. O projeto de Churchland exemplifica o problema que se encontrará: quando ele sugeriu o uso de estereótipos sedimentados sob o gênero das representações distribuídas, encontrou o desafio de fazê-los apresentar sistematicidade. Esses estereótipos (ou mesmo os obscuros “estilos”) são fruto de aprendizagem, inclusive a que se dá no âmbito da cultura e dos hábitos sociais. O projeto de Churchland, contudo, não conseguiu superar a dificuldade de construir redes conexionistas que apresentem a sistematicidade da cognição humana. Nos termos de Cherniak, isso significa que a criatura cognitiva imaginada por Churchland não seria capaz de atender a condição de consistência mínima, por exemplo. Afinal, tal como largamente discutido, seria necessário mostrar que uma criatura descrita naqueles termos tem condições de apreender as “regras certas” que permitem a agentes racionais satisfazer a condição de consistência mínima. Regras de consistência, mesmo mínimas, não podem ser fruto único de “estilos cognitivos” particulares porque, fosse o caso, haveria uma pluralidade de condições de consistência mínimas. Fosse assim, seria impossível qualquer tipo de predição comportamental, visto que a informação poderia ser arranjada de um modo completamente estranho e particular a cada criatura. Ao mesmo tempo, regras de consistência também não podem ser fruto

de elementos inatos, pois eles parecem sensíveis à enorme variedade de contextos socialmente estabelecidos em que o ser humano pode se encontrar. É verdade que, em várias ocasiões, Cherniak parece sugerir que estes diferentes tipos de mecanismos poderiam conviver, mas o que está em jogo é precisamente a formulação de uma teoria que explicite como isso é possível, e neste sentido, nenhum passo substancial foi dado.

Cherniak preocupa-se em estabelecer a conexão entre a racionalidade humana e aquilo que ele chama de compartimentalização (que, para todos os efeitos nesta investigação, é o mesmo que um estereótipo) do conhecimento, mas não vai além disso. Com efeito, seu ponto de chegada é o mesmo a que se chegou no primeiro capítulo deste estudo: a explicação mecanicista da inteligência humana tem uma íntima relação com o modo dinâmico com que a informação é organizada. Porém, como Cherniak não desenvolveu um modelo de como a informação deve ser organizada, nem detalhou suficientemente como tais mecanismos cognitivos operariam, pode-se concluir preliminarmente que ele não apresentou material suficiente para mostrar que o FP, em particular no seu aspecto organizacional, pode ser evitado. Cherniak destacou a importância da sensibilidade ao contexto ao atribuir racionalidade a uma criatura, mas não explicou como tal sensibilidade é possível, ao menos não de um modo imune ao FP.

Tal como em Pollock, é possível notar em Cherniak a tentativa de conceber algum tipo de espaço intermediário entre o contexto proximal e o contexto distal. No caso do primeiro, esta era o plano mestre. Já no último, isto é visível na tentativa de estabelecer uma memória de curto prazo com a qual o sistema cognitivo possa trabalhar. Ambos falham no mesmo momento, que é o de explicar como esse espaço intermediário se constitui a partir do contexto distal. Novamente, o resultado da estratégia de substituir a racionalidade plena por uma noção de racionalidade limitada parece enfatizar, e não resolver, a dependência que os processos cognitivos possuem em relação ao contexto distal. É inegável, contudo, que Cherniak desenvolveu um campo mais fértil que o de Pollock, no interior do qual teorias da cognição mais elaboradas poderiam ser desenvolvidas. Ainda não se pode descartar a possibilidade de que alguma versão de uma tal teoria pautada pela racionalidade mínima ou por alguma variante de racionalidade limitada possa mostrar-se sensível ao contexto distal e imune ao FP.

Há uma teoria que, sugere-se aqui, é bastante fiel ao que Cherniak defende em sua obra, que é a *teoria da relevância*[3], de SPERBER; WILSON (1995). Trata-se de uma sugestão local a esta investigação porque o enquadramento da TR como uma teoria da cognição guiada por uma racionalidade limitada pode ser um ponto de debate por si só. Há, contudo, uma forte razão para aceitá-la como tal: ainda que Sperber e Wilson não busquem discutir diretamente aspectos deônticos de racionalidade, eles clamam explicitamente (1996) que sua teoria é capaz de resgatar os processos racio-

[3] Daqui por diante, TR.

nais centrais da inexplicabilidade que Fodor lhes atribui. Este ponto, espera-se, ficará mais claro na medida em que a TR for aqui apresentada. O próximo passo, portanto, é descrever alguns aspectos centrais da teoria, para que se possa então apreciar seu desempenho face ao desafio da sensibilidade ao contexto distal e, consequentemente, ao FP.

3.2.3 A teoria da relevância

O que é a *teoria da relevância* (TR)? Apesar do nome, não se trata de uma teoria sobre o que é relevância, tal como esse conceito é utilizado pelo senso comum. A TR desenvolve uma noção particular de relevância, que é utilizada no interior da teoria. Embora por vezes os sentidos possam coincidir, não há compromisso forte entre o uso do termo no interior da teoria e o significado comum. Assim, para minimizar a chance de confusão, a relevância tal como compreendida no interior da TR será aqui designada *relevância-t* (o mesmo se fará com suas derivações, relevante-t, irrelevante-t etc).

O que significa então relevância-t e quais os papéis que Sperber e Wilson atribuem a este conceito? Antes de apresentar o modo como ele pode ser usado na construção de uma teoria cognitiva, é preciso compreender o conceito geral que o subjaz. A relevância-t é uma espécie de função custo-benefício. O benefício é a utilidade que a informação tem ao agente. O custo, por sua vez, é o esforço requerido, por parte dos mecanismos cognitivos, para processá-la. *Ceteris paribus*, quanto maior o efeito (isto é, a utilidade) de uma informação, maior sua relevância-t. *Ceteris paribus*, quanto maior o esforço necessário para processá-la, menor sua relevância-t. Assim, diante de informações como:[4]

(5) Você ganhou 500 reais.
(6) Você ganhou 10 reais, 500 reais ou 1000 reais.

Nesse caso, (5) é mais relevante-t que (6), visto implicar o mesmo que (6) a partir de um processo menos custoso. Para os autores, este princípio abstratamente simples pode ser útil de vários modos. Um dos modos mais conhecidos é o seu uso na caracterização do chamado *princípio da relevância* em discussões pragmatistas no âmbito da filosofia da linguagem. Tal princípio, argumenta-se, poderia substituir de modo eficaz as máximas conversacionais griceanas, tais como a Máxima da Quantidade (não informar nem mais nem menos do que é exigido), a Máxima da Maneira (clareza) etc.[5] Esta proposta fez com que a TR fosse amplamente discutida no âmbito da pragmática. Contudo, o uso que interessa à presente investigação é o que se faz no interior de uma teoria da cognição humana. Para Sperber e Wilson, a relevância-t deve ser uma característica fundamental de qualquer destas teorias. Para que se

[4] Exemplo adaptado de SPERBER; WILSON (1996).
[5] Conforme GRICE (1991).

possa descrever o papel da relevância-t que os autores querem em uma teoria cognitiva, é preciso antes apresentar alguns aspectos mais elementares.

Em *Relevance*, Sperber e Wilson esboçam os pilares fundamentais uma teoria cognitiva pautada pela relevância-t. Nela, os autores introduzem sua própria concepção de agrupamento informacional com caráter estereotípico. Estes são denominados *conceitos*. Conceitos, assim compreendidos, são aquilo por meio do qual a informação é compartimentalizada (para usar um termo de Cherniak) e abrangem tanto elementos lógicos, como no caso do conceito de *modus ponens*, quanto elementos lexicais, tal como no caso da relação entre um termo linguístico e seu significado. Há diferenças importantes: enquanto conceitos lógicos podem ser facilmente aceitos como inatos, dizer o mesmo de um elemento lexical seria bem mais controverso, por exemplo. Além disso, o termo é também usado para abranger entradas com informações de caráter enciclopédico, de modo equivalente ao que vem sendo referido aqui como *estereótipo*. Tal como em qualquer outra versão da TRM, crenças sobre o mundo são estruturadas em termos de informações armazenadas nestes estereótipos. Como a discussão estará centrada nos conceitos que abarcam tais informações estereotípicas, optou-se por manter o uso regular do termo "estereótipo" tal como se vem fazendo desde o início.[6]

Na TR, toda crença possui uma característica adicional, que é sua força. Qualquer aspecto do mundo que é tomado como verdadeiro se dá com um certo grau de força ou certeza. A força, no entanto, não é apenas uma medida com caráter epistêmico sobre quanta segurança um agente tem acerca de uma determinada crença. Embora, claro, a noção de força harmonize com esse aspecto epistêmico, ela tem um papel ainda mais fundamental na cognição, visto que é pressuposta uma conexão direta entre o quão forte e o quão acessível (isto é, o quão facilmente disponível) uma dada crença é. Intuitivamente, uma tal acessibilidade está conectada ao modo como a crença foi obtida, isto é, à confiabilidade da fonte. Trata-se, claro, de uma noção vaga e não necessariamente representacional de confiabilidade: compare-se, por exemplo, o grau de confiança que um agente tem diante do que ele vê em contraste com o grau de confiança que ele tem diante de um testemunho de alguém. Além disso, a força de uma crença pode ser continuamente revisada em função de reforços cognitivos. Se um amigo afirma gostar das músicas do *Pink Floyd* e isto é tomado como verdadeiro, então testemunhar sua ida a um show desta banda conta como uma "reafirmação" desta crença, tendo como efeito uma ampliação de sua força e, consequentemente, de sua acessibilidade.

Dado que, na TRM, toda crença é tratada representacionalmente, cabe então a pergunta: de que modo a força figura na teoria cognitiva de Sperber e Wilson? Se for tratada também representacionalmente, isto significa que cada crença no sistema

[6] Como é possível notar, trata-se de um uso muito específico da palavra "conceito". Talvez merecesse também ser denominado conceito-t, mas como a discussão não será focada neste aspecto da teoria, optou-se por tentar manter o texto menos poluído.

será, na verdade, um par crença-força. A força precisará ser então representada a partir de alguma escala métrica. Enquanto uma crença pode ter, por exemplo, uma força 95, outra poderia ter uma força 80 em uma escala de 1 a 100. Essa seria uma noção muito semelhante à utilizada por Pollock. No entanto, esta não é a proposta da TR:

> (...) estas variações na força não são nem objeto, nem resultado de uma computação lógica especial. Em vez disso, elas emergem como subprodutos de vários processos cognitivos, dedutivos e não dedutivos. (1995, p. 77)

Presumivelmente, a força é um elemento central na explicação do comportamento de qualquer agente. Assim, ao fazer da força algo não representacional, Sperber e Wilson terminam por violar o requisito da explicitude. Como já visto anteriormente, quando discutindo o conteúdo intencional implícito nos termos propostos por Cummins, isso não é necessariamente um problema para a TRM. Tudo depende do modo como estes elementos não representacionais são tratados na teoria cognitiva. Sperber e Wilson especulam que o esforço necessário para certos processos cognitivos é, de alguma forma, sedimentado de modo não representacional. Como se pode perceber, trata-se de uma noção conectada com o que Cherniak chamou de estilo cognitivo, inclusive no apelo a mecanismos que não fazem uso de processos racionais. Resta saber se tal concepção é suficiente para fazer deste um conceito efetivamente útil para tratar da sensibilidade ao contexto no âmbito de uma teoria da cognição. Embora este ponto vá ter importância no futuro, manter o foco nele seria despropositado, ao menos por ora. O importante no momento é manter em mente que todo o conjunto de crenças de um sujeito são armazenadas e processadas de modo sensível à força, isto é, à condições de acessibilidade não representacionais.

A força com que uma crença pode ser mantida é um aspecto fundamental dos mecanismos inferenciais concebidos nos termos da TR. Ao mesmo tempo, de modo geral, a teoria cognitiva esboçada em *Relevance* tem muito em comum com versões clássicas da TRM cognitivista. Como operam então os mecanismos inferenciais?

Neste ponto, vale notar que boa parte do desenvolvimento da TR se deu tendo como objetivo a explicação da atividade comunicativa, que envolve, mas não exaure, as capacidades cognitivas. Por isso, boa parte das situações exploradas pelos autores envolvem exemplos oriundos da comunicação. Como os mesmos colocaram, contudo, a TR é constituída por teses específicas acerca da comunicação, mas também por teses gerais sobre a cognição humana (1996). Esse movimento de trazer para o interior de uma teoria cognitiva aquilo que é discutido no âmbito da pragmática linguística pode ser visto, por exemplo, quando os autores discutem a natureza de processos centrais no âmbito da filosofia da mente. Ao falar da natureza inferencial do processo de compreensão linguística, Sperber e Wilson partem do princípio de que esse processo é isotrópico, tal como descrito por Fodor:

> Nós sustentamos que a compreensão inferencial não envolve mecanismos especializados. Em particular, vamos argumentar que a camada inferencial da compreensão verbal envolve a aplicação de processos inferenciais centrais não especializados sobre o resultado de processos linguísticos especializados e não inferenciais. (SPERBER; WILSON, 1995, p. 66)

A partir disso, os autores esboçam aquilo que constitui, de modo mais geral, os mecanismos inferenciais da cognição humana. Eles acreditam que esta é sim caracterizada por processos que não são informacionalmente encapsulados. De modo inicialmente surpreendente, eles retomam a tese de que os processos cognitivos tem caráter dedutivo, tal como se fazia nas abordagens pautadas pela logica clássica nos primórdios da GOFAI. Contudo, eles aparentemente reconhecem a ameaça do aspecto inferencial do FP, que Fodor denominou *problema de Hamlet*, isto é, em paráfrase, o problema de decidir quando parar de processar:

> . . . o sucesso das inferências não demonstrativas do ser humano precisa ser explicado não por apelo a processamento lógico que busque confirmar uma suposição, mas sim por restrições na formação e exploração de suposições. (1995, p. 81)

Essa afirmação mostra que eles estão cientes dos problemas relacionados ao uso da abordagem computacional clássica: é preciso fazer com que os processos "saibam" quando parar. Como exatamente operam estes mecanismos dedutivos concebidos na TR? Como de costume, eles atuam sobre as informações armazenadas no sistema cognitivo, isto é, sobre o conjunto de estereótipos que constitui o conhecimento que o sistema possui acerca do mundo.[7] No entanto, os processos inferenciais nunca se dão diretamente entre um *input* e a totalidade das informações armazenadas (o que Cherniak denominaria memória de longo prazo). Quando diante de um determinado *input*, que pode ser tanto oriundo da memória quanto da percepção, este tem seus efeitos determinados a partir de um subconjunto destas informações, análogo ao que Cherniak compreendia como a memória de curto prazo, e que será aqui denominado *contexto-t*.[8] Os mecanismos dedutivos atuam então sobre a conjunção do *input* com o conteúdo do contexto-t, buscando mapear a existência do que os autores denominam *efeitos contextuais*.

Na TR, efeitos contextuais são compreendidos e mapeados de diferentes formas. Um primeiro tipo de efeito, denominado *implicação contextual*, caracteriza uma informação nova, que não poderia ser derivada nem somente do *input*, nem somente do contexto-t, mas apenas da conjunção destes. Um segundo tipo de efeito contextual especialmente importante é o que envolve a noção de força. Tanto uma crença

[7] Nunca é demais relembrar que o termo "conhecimento" tem um significado especial no interior de uma teoria computacional, possuindo sentido distinto daquela que figura em teorias filosóficas no âmbito da epistemologia.

[8] O termo "contexto-t" é utilizado para referir-se ao modo distinto como a TR compreende o contexto e seu papel nos processos cognitivos, evitando assim possíveis confusões com outras noções de contexto, inclusive as de contexto proximal e contexto distal. Embora estas possam coincidir algumas vezes, é importante que o vocabulário adotado permita apontar quando este é ou não o caso.

representada no *input* quanto uma crença já disposta no contexto-t podem ter suas respectivas forças revisadas. A força com que um dado *input* é tomado pode afetar a força com que determinadas crenças presentes no contexto-t são tomadas, seja para reforçá-las, seja para mitigá-las. Isso significa que, de modo geral, os mecanismos dedutivos presentes no sistema cognitivo são sensíveis à força com que as crenças são armazenadas, e isto tem efeito sobre o quão acessíveis elas são.

Neste ponto, já é possível descrever o modo como a relevância-t é compreendida no interior deste sistema cognitivo: a relevância-t de um *input* é medida a partir dos efeitos contextuais que ela é capaz de gerar. Um *input* que não gera nenhum tipo de efeito contextual não possui relevância-t alguma, por exemplo. Dado, contudo, que revisões na força de uma crença estão entre os possíveis efeitos contextuais e, consequentemente, entre os possíveis elementos que podem determinar a relevância-t de um *input*, então um sistema que se paute pela relevância-t é, necessariamente, um sistema sensível à força com que as crenças são armazenadas. Como a força de uma crença não é tratada representacionalmente, não é possível computá-la. Gera-se, assim, o problema de explicar como mecanismos cognitivos de caráter computacional e racional podem demonstrar sensibilidade à relevância-t. Esta é uma dificuldade análoga à enfrentada por Cherniak quanto este fala em estilos cognitivos sedimentados de modo tal que não se pautam por qualquer processo racional. Sperber e Wilson apresentam algumas especulações a respeito:

> (...) efeitos contextuais e esforço mental, assim como movimentos físicos e esforço muscular, precisam causar alguma mudança físico-química sintomática. Podemos assumir que a mente avalia seus próprios esforços e seus efeitos ao monitorar essas mudanças. (SPERBER; WILSON, 1995, p. 131)

Infelizmente, este é um aspecto pouco desenvolvido na teoria, mas é possível tentar esclarecer a ideia dos autores por meio de uma analogia com outras formas de atividade: quando realizando qualquer atividade física, por exemplo, o corpo assume uma determinada configuração, fazendo uso de uma certa distribuição de energia entre as partes envolvidas na atividade. Assim, o corpo pode assumir um determinado perfil de distribuição de recursos entre os diferentes componentes do organismo quando sentando-se em uma cadeira, e outro perfil quando buscando superar um obstáculo na calçada, e assim por diante. Esse "perfil" de distribuição energética é o que os autores parecem ter em mente ao dizer que o modo como a informação é acessada é sensível à força com que a crença é armazenada. Nesse sentido, é como se o uso da informação envolvesse realizar diferentes acessos, percorrer diferentes estradas, cujas condições podem variar. Quanto mais força uma crença possui, menos caminhos precisam ser trilhados até que ela se torne acessível, e mais suave é a estrada. Crenças mais fracas são menos acessíveis e estão, portanto, sujeitas a caminhos tortuosos e estradas em más condições, demandando assim um maior esforço e um maior dispêndio de recursos cognitivos.

Um primeiro, mas fundamental, problema que salta aos olhos diante dessa especulação (para além do fato de ser uma especulação) é que ela não trata adequadamente daquilo que é demandado. Se o organismo tiver que realizar ou simular (representacionalmente ou não) uma dada inferência ou ação para que possa então tomá-la como relevante-t, nenhum avanço teria sido obtido. O organismo precisa, de algum modo, ser capaz de predizer a relevância-t, isto é, atuar com *expectativas de relevância-t*, o que faz lembrar de Pollock. No caso dele, trabalhou-se com a noção de expectativa de utilidade. A circunscrição prévia dada pelo plano mestre é o que permite que tal cálculo de expectativa seja finito e passível de realização por parte de seres com baixas capacidades cognitivas. Da mesma forma, a expectativa de relevância-t precisa ser tornada realizável por um organismo finito. Os autores sugerem dois possíveis caminhos, nenhum deles largamente desenvolvido, e que serão rapidamente apresentados.

O primeiro caminho é dado pela tese simples de que quanto maior a força de uma crença, maior a expectativa de relevância-t. Assim, supor que uma crença é tão mais acessível quanto mais forte, é o mesmo que supor uma organização informacional pautada por graus de acessibilidade. A partir disso, é possível supor que o organismo disponha de um subconjunto previamente circunscrito constituído pelas crenças mais fortes (consequentemente, as mais imediatamente acessíveis). A pergunta pelo quão forte uma crença deve ser para que possa participar do subconjunto poderia ser dada por algum parâmetro evolutivamente sedimentado (tal como Cherniak também especulou). O problema, claro, é que a organização seria excessivamente rígida. Sendo tal organização “fixa” em função da força da crença, disto se segue que todo e qualquer *input* é processado a partir da mesma informação, organizada da mesma forma. Isso gera diversas dificuldades já familiares: o que fazer com crenças que raramente são relevantes, mas que certamente ocupariam o subconjunto das crenças mais fortes, tais como aquelas acerca do nome próprio, do local de nascimento, da roupa que se está a usar, etc. Diante de um tocar da campainha, antes de levar em conta as possíveis reações, deve o organismo necessariamente considerá-las?

Embora a TR conceba ao *input* o poder de alterar essa organização, seja agregando crenças, seja alterando a força de uma ou mais delas, supor que isto seja feito sempre a partir de um mesmo critério organizacional leva ao FP. Isso porque, ainda que um *input* possa realizar efeitos sobre o conjunto de crenças, o organismo deve decidir se e quando esse é o caso. Com efeito, essa discussão sugere que uma tal ideia só funcionaria num ser finito se pressuposto um recorte adequado na informação considerada, isto é, um recorte prévio no interior do contexto distal. Algo como um plano mestre de Pollock, ou da memória de curto prazo de Cherniak. Nos termos da TR: esta hipótese só é factível se suposto um contexto-t previamente determinado.

Uma segunda tentativa, brevemente citada pelos autores, é o uso de contraste entre pares de possibilidades. Um volume de *input* maior, como uma sentença grande e complicada, por exemplo, presumivelmente demandaria um esforço maior em con-

traste com sentenças mais simples, justificando assim um “palpite” cognitivo acerca de qual caminho inferencial apresentará menor expectativa relevância-t. Contudo, esta é uma ideia que não parece aplicável a todo caso, em todo e qualquer *input*, mas somente onde este tipo de contraste se mostrasse necessário. Além disso, para que possa ser realizada por seres finitos, deve haver um número finito de testes de contraste a realizar. Novamente, a adoção de uma regra fixa parece levar à presunção de alguma forma de circunscrição prévia do contexto-t.

Como se viu, ambos os caminhos sugeridos apresentam um mesmo problema, que é a pressuposição de um contexto-t com conteúdo previamente determinado e estruturado. Não por acaso, este é precisamente o desafio que nem Pollock, nem Cherniak conseguiram superar. Enquanto o primeiro falhou em dar uma explicação plausível de como o plano mestre poderia ser adequadamente mantido sem que este processo recaísse no FP, o segundo apenas especulou sobre possíveis formas de manter a memória de curto prazo com o conteúdo e estrutura adequadas. Essa “pressão” que estas hipóteses sofrem está relacionada com as motivações típicas para rejeição da racionalidade plena. Afinal, se for necessário acessar toda informação presente no sistema cognitivo para determinar a relevância-t de uma inferência ou processo análogo, então esta seria uma propriedade global, nos moldes descritos por Fodor, sofrendo todas as consequências disto. Contudo, há razões para crer que os autores estão cientes desse problema. Resta-lhes, talvez, reconhecer o quão grave este é. Tome-se como exemplo o seguinte trecho em que os autores tratam da dificuldade no âmbito da comunicação:

> Se o contexto-t incluísse o todo da enciclopédia do ouvinte, virtualmente qualquer nova informação que um falante pudesse expressar seria relevante, visto que virtualmente toda nova informação teria algum efeito contextual em um contexto tão enorme. Por outro lado, dado o tamanho de tal contexto-t, um esforço de processamento enorme - para não falar do tempo de processamento - seria necessário para realizar estes efeitos. Como a relevância-t é reduzida na medida em que mais esforço é requerido, isto significaria que, ainda que qualquer nova informação atingiria relevância-t facilmente, nenhuma informação iria jamais alcançar mais que um grau mínimo de relevância-t. (SPERBER; WILSON, 1995, p. 137)

Com efeito, os autores percebem o problema de considerar o contexto-t como sendo equivalente ao contexto distal. Tome-se como ilustração disso o seguinte trecho:

> (. . .) nem todos os blocos de informação enciclopédica são igualmente acessíveis em uma dada ocasião. Poderia ser, por exemplo, que a entrada enciclopédica de um conceito torne-se acessível somente quando aquele conceito [estereótipo] aparece em alguma assunção que já tenha sido acessada. (. . .) Haverá ocasiões em que esta informação será acessível em um único passo, ocasiões em que ela estará acessível a partir de diversos passos, cada um envolvendo uma extensão do contexto-t, e ocasiões em que o número de passos envolvido irá, na prática, tornar a informação inacessível. (1995, p. 138)

Nesse caso, a relevância-t se mostra uma meta-dependência contextual, isto é, depende de uma consideração prévia do contexto distal em sua totalidade. Sendo este o caso, retornam problemas já familiares e discutidos: como realizar, por exemplo, a migração de um estereótipo a outro em função de um *input*? Como detectar casos em que um *input*, ou um conjunto deles, devem ser interpretados sob um dado contexto proximal, e não sob outro? Em síntese: como demonstrar sensibilidade ao contexto distal? Como se viu, fazer da força, ou da própria relevância-t, um elemento não representacional, não exime a TR de explicar como esses elementos não representacionais operam sobre ou afetam entidades representacionais de modo a fazer com que o FP deixe de ser uma ameaça. Parte dessa explicação envolve, necessariamente, uma descrição dos mecanismos por meio dos quais o contexto-t é constituído, isto é, de como este tem seu conteúdo selecionado e estruturado. Somente quando este desafio for vencido, será possível fazer uma defesa contundente da tese de que todo *input* é processado tendo como pano de fundo o contexto-t, e não a totalidade das informações presentes no sistema cognitivo. Este ponto, aliás, caracteriza a recusa dos autores em pautar-se pela racionalidade plena: não é necessário, como supõe Fodor, trazer à tona todo o conjunto informacional ao processar um *input*. Basta o conteúdo do contexto-t, tomado este como um conjunto de informações cuja organização é sensível à força de cada uma delas. Isso é o que lhes permite negar, contra Fodor, a inexplicabilidade de mecanismos racionais.

Mas como, afinal, constitui-se o contexto-t? Sem surpresa, Sperber e Wilson não oferecem uma tese amadurecida a esse respeito, mas somente algumas ideias tomadas como plausíveis. Uma síntese dessas ideias mostra que tanto o caráter da solução, quanto o caráter dos problemas enfrentados, são análogos aos de Pollock. Também na TR, o conteúdo do contexto-t é modificado de modo incremental, a partir dos *inputs*, trazendo à tona a pergunta: quais *inputs*, e sob quais circunstâncias, devem gerar modificações no contexto-t? Nos termos dos autores, este é apenas outro modo de perguntar como determinar qual o processo inferencial com a maior expectativa de relevância-t. Em termos ainda mais específicos à TR, este é um modo de perguntar qual o caminho inferencial capaz de gerar o maior número de efeitos contextuais a partir da conjunção entre *input* e contexto-t.

Essencialmente, supõe-se que, a cada *input*, o contexto-t seja afetado, podendo ser expandido, ter informações reduzidas ou reorganizadas (por meio de efeitos contextuais que afetem a força das informações ali contidas). O resultado é um contexto-t que se mantém como um conjunto de contextos proximais que orbitam adequadamente os *inputs*. O contexto-t caracteriza, nesse sentido, a "afinação" do agente às suas circunstâncias. Essa afinação precisa dar conta de dois momentos importantes da vida cognitiva: a migração entre contextos proximais e a decisão sobre quando e como rever o conteúdo do contexto-t. Primeiro, como se dá a escolha do contexto proximal (neste caso, do estereótipo) adequado dentre aqueles disponibilizados pelo contexto-t? Na TR, este se dá pela expectativa de relevância-t. Partindo-se

de um conjunto previamente circunscrito, utiliza-se um processo norteado pela busca por relevância-t para que se selecione o contexto proximal mais relevante nas atuais circunstâncias. Note-se, contudo, como isso pressupõe um contexto-t organizado em termos de acessibilidade, isto é, da força pertinente a cada crença. Sem isso, o uso da relevância-t como critério falharia. O contexto-t não é apenas um conjunto razoável de contextos proximais, mas um conjunto *apropriadamente estruturado* destes. É isso o que os permite afirmar que *"... evidências relevantes são mais provavelmente encontradas seguindo um caminho de menor esforço"* (1996, p. 532).

Isso leva ao segundo momento: tal como discutido desde o início, sempre haverá a circunstância em que o agente precisa migrar de um contexto proximal ao outro de modo mais radical. Essa mudança pode envolver a busca por informações estereotípicas adicionais não circunscritas ao contexto-t, mas presentes no todo das informações contidas no sistema (na memória de longo prazo, diria Cherniak). O agente pode se perceber não mais num restaurante, mas sim num hospital, por exemplo. Os estereótipos adequados devem ser então acrescidos ao contexto-t, e isto pode se dar de modo concomitante ou não a outras modificações. Isso traz à tona um velho problema: se o *input* for sempre tratado à luz do conteúdo do contexto-t atual, então deverá ser possível sempre, sob a luz dos estereótipos ali contidos (como o de um restaurante, por exemplo), determinar que é o momento adequado de migrar para uma outra versão do contexto-t (como o de um hospital). Isso significa que, dentre as informações estereotípicas do restaurante, devem constar orientações específicas sobre como migrar para um contexto proximal hospitalar (ou esportivo, ou profissional, ou uma versão nuançada destes, etc). Ora, se algo assim for necessário para cada possível contexto proximal, o que se tem é uma nova formulação do FP, afinal, o processo que mantém o contexto-t atualizado em função dos *inputs* não consegue tratar adequadamente o contexto distal: o agente está diante de uma meta-dependência contextual. Em alguma medida, os autores se mostram cientes desse problema:

> ... não é suficiente dizer que a informação pode ser carregada de um processo conceitual para o próximo; é preciso saber qual informação é mantida na memória de curto-prazo, qual é transferida para a memória enciclopédica e qual é simplesmente descartada. (1995, p. 138–139)

No caso de Cherniak, foi num momento análogo a este que ele apelou para noções vagas de estilos cognitivos e mecanismos evolutivamente sedimentados. Da mesma forma, Sperber e Wilson não ofereceram, ainda, nenhuma hipótese amadurecida, mas somente algumas especulações. Não é possível desenvolvê-las adequadamente aqui porque os próprios autores não o fazem, nem mesmo em textos posteriores.[9] Resta, portanto, apenas mencioná-los: um primeiro modo sugerido seria o de "voltar atrás" e acessar a informação utilizada para *inputs* anteriores. A ideia, essencialmente, é a de que *inputs* pretéritos que não geraram efeitos contextuais podem vir a

[9] Ver SPERBER (2005), por exemplo.

gerá-los em função de um novo contexto-t, contra o qual estes são agora processados. Isso soa algo bastante semelhante com a sugestão de que haveria uma espécie de contexto-t de segundo nível, que armazenaria mesmo informações irrelevantes-t, na expectativa de que possam vir a ser úteis no futuro próximo (trata-se, portanto, de uma estratégia heurística). A pergunta sobre até que ponto seria possível voltar atrás seria tratada empiricamente, a partir de pesquisas sobre a capacidade cognitiva humana. Um segundo modo de incrementar o contexto-t seria atribuindo um papel de destaque aos dados dos sentidos: o ambiente físico e os arredores seriam continuamente incorporados ao contexto-t, afetando-o, talvez, com um grau maior de relevância-t pressuposta (uma espécie de viés inato para certas fontes de *inputs*). Um terceiro modo seria acrescer ao contexto-t tudo aquilo que estiver contido nos estereótipos de quaisquer dos elementos das crenças envolvidas. Se um determinado *input* envolve, por exemplo, jogos de azar, então os estereótipos envolvendo jogos de azar passariam a fazer parte do contexto-t. Trata-se de um incremento "temático". É por meio de mecanismos deste tipo, acreditam Sperber e Wilson, que os agentes conseguem manter um contexto-t constantemente adequado (ainda que não infalível) e afinado com o mundo ao seu redor, incluindo aí os aspectos físicos do ambiente e as situações sociais.

Há problemas específicos a cada uma das sugestões e um problema mais geral. No caso da terceira sugestão, por exemplo, segundo a qual uma informação é acrescida ao contexto-t em função de alguma porção do *input* que a referencia, um *input* que remeta a cavalos faria com que os estereótipos relacionados a cavalos fossem trazidos para o interior do contexto-t. O problema é que esse mecanismo de tratamento da memória precisa apresentar sensibilidade ao conteúdo. Como se viu, este é o tipo de propriedade apresentada por abordagens conexionistas pautadas pelo uso de representações distribuídas. Porém, isso só é possível ao custo da sistematicidade, algo de que a TR não pode abrir mão, tornando esta uma sugestão pouco significativa. Como os autores não desenvolveram a hipótese, é difícil dizer se aceitariam este ponto como um problema fatal ou não. A segunda sugestão, por sua vez, possui limitações evidentes. Um barulho agudo ou a aproximação rápida de um objeto podem constituir *inputs* mais relevantes-t em função de um viés evolutivamente sedimentado, mas que podem permanecer irrelevantes-t, a depender das circunstâncias, e este aspecto não seria tratado adequadamente por tal mecanismo. Por fim, a primeira sugestão é representativa de um problema mais geral. Este deverá ser resolvido pelo mecanismo, ou pelo conjunto de mecanismos que explica o modo como o contexto-t é constituído e mantido, seja ele qual for: a dependência da relevância-t mesmo quando lidando com informações externas ao contexto-t.

Como se viu no caso de Pollock, este falhou em fornecer um critério adequado para tratar do problema de como delimitar os candidatos a mudanças no seu plano mestre. Nos termos de Cherniak, trata-se do desafio de decifrar a compartimentalização da memória de longo prazo, pois esta permitiria uma "navegação" adequada das informações. Na TR, por sua vez, existe o desafio análogo de como lidar com o

conteúdo não presente no contexto-t.

Nesse momento, Sperber e Wilson fazem uso de um modelo análogo ao da comunicação. Em casos típicos, o ambiente cognitivo compartilhado por falante e ouvinte modula aquilo em função de que a compreensão se dá, isto é, constitui o pano de fundo contra o qual os enunciados são avaliados. O que permite que tal modelo seja funcional no âmbito da comunicação, como já visto, é que ele possui um norte descritível em termos da relação professor-aluno, tal como apresentada no capítulo anterior. Professor e aluno agem seguindo uma mesma orientação, e isso permite a atividade comunicativa. O aluno consegue manter seus processos cognitivos no interior de um contexto-t gerado e mantido a partir desse norte, considerando irrelevante tudo aquilo que está fora dele durante a ação. Sperber e Wilson acreditam que a cognição humana, mesmo quando não restrita à comunicação, possui estrutura análoga:

> As habilidades cognitivas do ser humano são parte da natureza; eles são bem adaptados como resultado de evolução natural. Pode ser o caso que, das assunções [crenças] que vem mais espontaneamente à uma mente humana, aquelas que são verdadeiras são as mais provavelmente relevantes-t do que aquelas que são falsas, de modo que quando a relevância-t é alcançada, ela provê reforço retroativo geralmente válido. Se este é o caso, então a sugestão de Fodor, de que o pensamento científico pode ser tomado como típico de processos de pensamento centrais está errada. A natureza ajuda seres humanos a desenvolver uma genuína, ainda que limitada, compreensão dela (. . .). (1995, p. 117)

Com efeito, os autores rejeitam o modelo da empreitada científica como característico da racionalidade, substituindo-o por um modelo de racionalidade limitada inspirado na atividade comunicativa. A racionalidade plena é uma abstração da racionalidade: ao destacá-la da natureza, ela perde seu rumo, e este rumo é o que permite ao organismo evitar o FP ao “escolher” o conteúdo do contexto-t. A racionalidade humana é guiada: ela evoluiu assim para tal. Isso a torna falha, mas também a torna possível para criaturas com recursos finitos. Com efeito, na TR, o papel de circunscrever e manter um contexto-t adequado é, ele mesmo, guiado. Mas que norte seria esse? O que poderia funcionar como o “professor” da cognição em geral? Sperber e Wilson não possuem uma teoria bem desenvolvida a respeito disso, mas especulam que, como a noção de força com que a crença é tomada tem paridade com a noção de acessibilidade, esta deve também ser epistemicamente guiada. Assim, o norte do aparato cognitivo humano é o de incrementar o conhecimento individual do mundo: *“Isso significa adicionar mais informação, informação que é mais acurada, mais facilmente acessível e mais desenvolvida em áreas de maior interesse ao indivíduo.”*. (SPERBER; WILSON, 1995, p. 47) Esse norte inato e evolutivamente sedimentado pautaria toda a vida cognitiva do agente, afetando o modo como todo *input* é tratado, seja na hora de armazená-lo na memória de longo prazo, seja na hora de coletá-lo para que constitua o contexto-t e mesmo para guiá-lo no interior do contexto-t. A ideia de uma cognição assim norteada tem um efeito positivo, que é o de tornar mais inteligível a noção de

“estilo cognitivo”, citada por Cherniak: o conteúdo exato dos estereótipos que constituem o conhecimento do agente sobre o mundo pode variar em função do acúmulo de experiências, mas são sedimentados e processados sempre sob um norte fixo. Em princípio, esse cenário permitiria explicar em função de que o sistema consegue manter consistência e flexibilidade sem ceder ao FP. Resta saber se este é, de fato, o caso.

Como se viu até aqui, a TR fornece um quadro teórico bastante elaborado, mesmo que caracterize apenas um esboço de uma teoria cognitiva. Pode-se agora avaliar o que fora apresentado de modo mais diretamente ligado ao fio condutor utilizado nesta investigação: pode a TR fornecer um tratamento adequado das meta-dependências contextuais? Pode a TR explicar a sensibilidade ao contexto distal? Pode o FP ser evitado por apelo a um modelo cognitivo que rejeite a racionalidade plena e que seja pautado por um norte evolutivamente sedimentado? Chiappe e Vervaeke (1997) elaboraram um desafio à TR, afirmando que o modo como o contexto-t é tratado não é capaz de evitar o FP, e que este seria um problema fundamental para a TR.

Em sua resposta Sperber e Wilson (1996) defenderam, essencialmente, que o critério de racionalidade plena proposto por Fodor deveria ser rejeitado. Tal rejeição se concretizaria no fato de que os processos cognitivos são pautados pelas informações presentes no contexto-t, e não na totalidade de informações existentes no sistema cognitivo. Desse modo, as decisões tomadas e as ações realizadas são sempre pautadas pelo que parece satisfatório ao agente nas circunstâncias dadas pelo contexto-t, não sendo necessário acesso a qualquer propriedade global, como exigido no modelo de Fodor. O problema com essa resposta, claro, é que ela supõe um contexto-t previamente circunscrito e estruturado de modo a caracterizar a afinação do agente com o ambiente. Cientes disso, Sperber e Wilson tentam resolver esse problema a partir de sugestões e especulações sobre o tipo de mecanismo que poderia ser responsável pela manutenção do contexto-t. Estes mecanismos seriam eficazes porque seriam guiados pela natureza, isto é, por mecanismos que foram sedimentados evolutivamente e que se pautam pelos “interesses epistêmicos” do agente na hora de organizar a informação sedimentada. E ainda, o fato de estes mecanismos não se pautarem exclusivamente por representações (como no caso da força com que cada crença é tomada), tornaria a TR imune a problemas computacionais como o FP.

Essa estratégia pode ser criticada. O apelo a noções não representacionais guiadas (força da crença e, por conseguinte, a própria relevância-t) não ajuda a TR evitar o FP. Este é o caso porque a própria métrica da força utilizada é uma meta-dependência contextual e, como tal, varia em função do contexto distal. Trocar um dólar por quatro reais pode ser um bom negócio em uma determinada circunstância, mas um péssimo negócio em outra. O mesmo tipo de variação circunstancial pode fazer com que valha a pena trilhar um determinado caminho inferencial até certo ponto ou não. Trata-se de um problema já desenvolvido no primeiro capítulo: a própria orga-

nização informacional é dependente das circunstâncias. Sendo este o caso, a adoção de um norte fixo não pode resolver o problema, pois ela geraria, e dependeria de, uma organização também fixa. O próprio norte a ser seguido, portanto, é uma meta-dependência contextual. Em alguma medida, Sperber e Wilson mostram-se cientes da dificuldade:

> É extremamente improvável que a importância relativa do efeito e do esforço permaneçam constantes em todas as circunstâncias e indivíduos. Por exemplo, variações no grau de vigilância [alerta] podem muito bem alterar a disposição em levar adiante um certo esforço de processamento: algumas vezes, a esperança de atingir um dado nível de efeito contextual vai ser suficiente, em outras, não. (1995, p. 131)

Apesar disso, eles não desenvolvem, para além do que foi aqui apresentado, os mecanismos por meio dos quais isto seja realizado pelo sistema cognitivo. Acreditam ser suficiente mostrar que a teoria é compatível com esta relativização da relevância-t com que uma crença ou processo pode vir a aparecer em função das circunstâncias (isto é, em função do contexto distal). Para os autores, isso se dá em função do caráter não representacional dessa organização: a acessibilidade não é fruto de aspectos representacionais das informações. Sendo assim, a descrição destes mecanismos extrapolaria o escopo de uma teoria computacional da cognição.

Um problema com o modo como os autores encaram esta questão é que eles parecem colocar peso excessivo sobre os estados “internos” do indivíduo: seu cansaço, seus desejos, seu ânimo etc. Contudo, uma sensibilidade legítima ao contexto distal precisa envolver pesos fornecidos pelas circunstâncias “externas” ao agente. Quão a sério deve ser levada uma afirmação de outrem? Os estados internos do agente tem evidente influência na resposta, e fazem parte do contexto distal, mas as circunstâncias “externas” tem uma influência muito mais frequente e pesada.

Casos extremos podem constituir alguns exemplos interessantes: quão a sério deve-se levar o que diz alguém armado? Quão a sério deve-se levar o que diz alguém que depõe a seu próprio favor em um tribunal? Estas são meta-dependências contextuais que alteram não apenas a força com que os *inputs* são tomados, mas também a força com que outras informações foram previamente sedimentadas no agente (os conjuntos de estereótipos conhecidos). Com efeito, se a relevância-t está sempre afinada ao grau de esforço necessário para realizar uma determinada operação, é preciso mostrar como ela se sensibiliza ao modo como o esforço necessário varia em função das circunstâncias. Por mais força que uma determinada crença tenha, as circunstâncias podem fazer com que ela “desapareça” do ambiente cognitivo, isto é, do contexto-t.

O cenário é familiar: assim como é necessária uma organização prévia em uma biblioteca para que uma busca por ordem alfabética seja eficiente, para que a relevância-t sirva como critério ao organismo (seja na realização de operações inferenciais sobre representações, seja por apelo a mecanismos não representacionais),

é preciso supor que a informação existente no sistema cognitivo do agente tenha sido previamente arranjada de modo adequado. Presumivelmente, é necessário mostrar como as circunstâncias alteram a estrutura das informações, estejam ou não dispostas no contexto-t. O que a TR precisa, portanto, é descrever algum tipo de mecanismo de afinação com o ambiente que seja, ao mesmo tempo, capaz de se reorganizar rapidamente em função das circunstâncias, mas sem abrir mão do caráter computacional da cognição humana. O norte do organismo deve ser, ele mesmo, dinâmico.

Como se viu no capítulo anterior, esta é a gangorra encontrada por Churchland ao propor o uso de representações distribuídas: ganha-se flexibilidade, mas perde-se sistematicidade. A TR vai um pouco além de de Churchland ao defender que representação alguma é necessária para sedimentar a força com que as crenças são tomadas, mas não vai além de Cherniak ao especificar como isso seria possível.

Sperber e Wilson poderiam ainda insistir e argumentar que a relevância-t (bem como a força de uma crença) não precisa ser representada, mas que ela pode ser. Tome-se o exemplo apresentado há pouco: quão a sério levar aquilo que é dito por alguém armado ou por alguém que advoga em causa própria? Isso pode sim, diriam os autores, ser tratado em termos de crenças representadas no interior do sistema cognitivo do agente. Ele pode dispor da crença de que está diante de alguém armado, e de que pessoas armadas são perigosas se contrariadas, e assim por diante, constituindo então um modo representacional de afetar a força com que a crença deve ser tomada. Para os autores, isso não leva ao FP porque o agente não trabalha com a totalidade das informações presentes no sistema cognitivo, nem tampouco com todas as possíveis derivações inferenciais delas, mas somente com aquilo que está presente no contexto-t. Porém, esta linha de argumentação parece trazer sérios prejuízos: além de não satisfazer a dependência de uma explicação adequada sobre como esse contexto-t é formado e mantido, ela traz consigo duas dificuldades adicionais: primeiro, ela faz deste um problema que não pode ser delegado à teorias outras. Por serem representacionais, os mecanismos envolvidos fazem parte do que é abarcado por uma teoria computacional. Segundo, torna-se necessário fornecer uma descrição adequada de como e quando o agente detecta as circunstâncias em que deve ou não representar a força de uma crença. Uma vez que, em alguma medida, os autores se mostraram cientes desse tipo de dificuldade, o ponto que talvez lhes tenha escapado não é o de que isso precisa ser explicado, mas sim o da magnitude do desafio e do quanto a TR depende da superação dele.

Considerando o volume de informações apresentadas nesta sessão, realiza-se agora uma pequena recapitulação e síntese da discussão feita sobre a TR. Como se viu, a teoria supõe a existência de um contexto-t que é análogo ao plano mestre de Pollock e à memória de curto prazo de Cherniak. Esse contexto-t pode ser expandido pelo acréscimo de informações, bem como reduzido pela remoção de outras, mantendo assim sempre um tamanho adequado e em conformidade com os recursos cognitivos do agente. O contexto-t é organizado em função da força das crenças que ali existem, e

essa força não é necessariamente representada. Isso significa que a acessibilidade da informação no interior do contexto-t pode se dar em função de mecanismos não representacionais que, como pontuou Cherniak, não necessariamente se pautam pela realização de processos racionais. Tendo essa organização pressuposta, o sistema pode fazer uso da relevância-t para "navegar" pelas informações no interior do contexto-t. A harmonia entre o contexto-t e a relevância-t se explica por apelo à natureza, que fornece um norte: ambos são fruto concomitante de um mesmo processo evolutivo de caráter epistêmico que busca ampliar ao máximo, e com o máximo de eficiência, o conhecimento do agente sobre seu ambiente cognitivo. Explica-se assim o motivo de a força com que uma crença é tomada estar relacionada com sua acessibilidade (ou seja, com o modo como ela se organiza), bem como a relação desta acessibilidade com a verdade ou adequação da crença às circunstâncias concretas em que o agente se encontra.

Contudo, é preciso dar conta também de explicar como o contexto-t é constituído. Em função de que uma informação pode figurar ou deixar de aparecer no interior do contexto-t? Os autores sugerem que isso pode se dar por uma pluralidade de mecanismos: "voltar atrás" e priorizar informações recentemente consideradas, priorizar dados dos sentidos, buscar por informações de modo temático com base no conteúdo dos *inputs* etc. No entanto, ainda que plurais, deve haver um ponto em comum a estes mecanismos: eles precisam ser guiados pela relevância-t. Sem isso, torna-se misterioso o modo como o contexto-t é constituído. Isso significa que o modo como a informação é acomodada no interior do contexto-t precisa refletir o modo como ela se organiza no exterior do contexto-t, isto é, na memória de longo prazo. Se uma dada crença da memória de longo prazo foi sedimentada com uma determinada força, então essa força precisa ser mantida quando ela passa a figurar no contexto-t. As relações estruturais entre os estereótipos são um exemplo: se a escolha do que beber apresenta tipicamente uma alta relevância-t no interior de um estereótipo de *estar-em-um-restaurante*, isto precisa ser mantido quando o estereótipo é "convocado" para figurar na memória de curto prazo.

Isso significa que a estrutura do contexto-t será, portanto, sempre derivada do modo como a informação é armazenada na memória de longo prazo. Ora, como se viu há pouco, a força com que cada crença ou a relevância-t de cada empreitada inferencial variam, elas mesmas, em função das circunstâncias. Trata-se de uma meta-dependência contextual. Isso significa que tanto a estrutura do contexto-t, quanto a estrutura da memória de longo prazo, precisam apresentar sensibilidade a estas circunstâncias. O contexto distal pode exigir do sistema cognitivo que ele reorganize as informações do contexto-t, mas os mecanismos que tratam desta reorganização dependem, eles mesmos, de que as informações contidas no restante do sistema cognitivo sejam reorganizadas em função das circunstâncias. Os processos que determinam o que irá constar no contexto-t, e sob qual arranjo, precisam ser, eles mesmos, sensíveis ao contexto distal. Cria-se assim, uma tensão: abrir mão da sensibilidade ao

contexto distal é o que permite o uso de um norte fixo (a busca constante pela melhor relevância-t possível); ao mesmo tempo, abrir mão de uma orientação fixa faz com que o sistema cognitivo perca o rumo e sofra do FP. Nesse sentido, este aspecto da TR é, diria Fodor, mais uma formulação do FP. Portanto, a TR não evita o surgimento do FP, mas antes pressupõe uma solução a ele. Sem isso, ela é incapaz de apresentar a necessária sensibilidade ao contexto.

3.3 Considerações finais

Antes de concluir, convém recapitular rapidamente o que foi discutido acerca da racionalidade. No começo do capítulo, foram feitas algumas considerações sobre o que seria uma concepção racionalista da inteligência humana. Essa concepção permearia tanto a inteligência tal como trabalhada no interior da GOFAI, quanto a maior parte das teorias da mente no âmbito da filosofia. Neste cenário, desdobrou-se a possibilidade de utilizar concepções alternativas de racionalidade para tratar da inteligência humana e, concomitantemente, para formular uma teoria da cognição adequada. Para que este debate pudesse se manter dentro da órbita adequada à presente investigação, formulou-se uma tese específica (T5) a ser perseguida. O objetivo geral era investigar se o abandono da racionalidade plena em virtude de diferentes concepções de racionalidade limitada poderiam afetar teorias da cognição de um modo tal que permitisse a superação do FP e a explicação da sensibilidade ao contexto distal.

Num primeiro passo sugeriu-se que a situação encontrada em Pollock constituiria uma espécie de padrão no modo como diferentes concepções de racionalidade atuariam sobre teorias da cognição: a adoção de uma racionalidade pautada pelas limitações cognitivas do agente se mostraria no estabelecimento de um campo intermediário entre o contexto distal e o contexto proximal. Tal espaço intermediário seria a contraparte cognitiva do abandono da necessidade de satisfazer condições ideais (como o acesso a propriedades globais fodorianas) em função de condições de satisfação mitigadas.

Em Pollock, isso ficou nítido na sua tentativa de formular o *plano mestre* como originado do acúmulo gradativo de planos locais. Este seria construído paulatinamente, na medida em que as experiências do agente acontecessem, não havendo, portanto, "recorte" algum a fazer em um plano universal, espaço global ou qualquer concepção análoga. A falha de Pollock estava, contudo, em não conseguir apresentar um modo de o agente constituir e manter um plano mestre sem o envolvimento de qualquer processo universal. Em particular, é necessário selecionar, dentre todos os *inputs* e dentre todas as possíveis inferências por eles permitidas, quais podem ser considerados candidatos em potencial para realizar alterações no plano mestre. Pollock tratou este problema como sendo um problema de implementação, ou seja, seria uma questão de identificar e descrever os mecanismos por meio dos quais isto se dá. Como se

viu, algo análogo se mostrou para Cherniak e para a TR.

No caso de Cherniak, o campo intermediário é a própria memória de curto prazo. A condição de racionalidade mínima, constituída pelas condições mínimas de inferência e consistência seriam respectivamente explicadas por teorias auxiliares. No caso, uma teoria das inferências viáveis e uma teoria da organização da memória. Ambas as teorias propostas mostraram-se dependentes de uma constituição adequada da memória de curto prazo, constituição esta que é pautada pela própria relevância. O modo como a memória de curto prazo é preenchida e organizada depende da sensibilidade àquilo que é relevante em cada circunstância específica. Como espera-se ter deixado claro, contudo, essa relevância termina sendo não explicada, mas pressuposta pelas teorias auxiliares de Cherniak. Cherniak não apresenta a ingenuidade de Pollock em supor que este seja um mero problema de implementação, mas acredita que a relevância deve ser explicada por apelo a mecanismos outros que seriam fruto de teorias que ele meso não desenvolve.

O mesmo padrão pode ser encontrado no caso da TR. Sperber e Wilson não se comprometem com nenhuma noção específica de racionalidade limitada, mantendo isto em aberto, mas deixando claro que rejeitam a racionalidade plena nos termos de Fodor. Para estes autores, agente algum considera todas as possibilidades diante dos *inputs*, nem precisa ter acesso a propriedades globais ou nada nestes moldes. Isso se viabiliza porque a cognição é norteada pela relevância-t. O modo como ela armazena e processa informações leva em conta a tese de que a relação do agente com o mundo se dá por mecanismos que foram evolutivamente sedimentados de modo tal que, o menor esforço coincide com a maior probabilidade da informação ser útil nas circunstâncias em que o agente se encontra. Os mecanismos responsáveis por isso constituem um espaço intermediário entre o contexto proximal e o contexto distal, denominado contexto-t. Como se viu, a TR vai além de Cherniak ao dizer que os mecanismos capazes de explicar o modo como o ser humano detecta aquilo que é relevante dentre os *inputs* está ligado a mecanismos cognitivos evolutivamente sedimentados e não necessariamente representacionais. Contudo, o ponto fraco da TR aparece no mesmo ponto em que as demais teorias são consideradas: como constituir e manter um contexto-t adequado? A TR demanda que tanto os mecanismos que constituem o contexto-t, quanto os que tratam de informações fora deste espaço sejam guiados pelo mesmo norte.

O problema é que o uso de um guia fixo e evolutivamente sedimentado não parece capaz de explicar a flexibilidade com que o ser humano lida com as circunstâncias ao seu redor. Ele pode servir para explicar a relevância-t atribuída a processos envolvendo situações mais simples, como a de um som agudo em contraste com um ruído contínuo, mas há um longo, e talvez intransponível, caminho entre isso e explicar como o ser humano apresenta comportamento adequado diante de um tocar da campainha. Numa situação como essa, há uma série de possíveis reações que, a depender do contexto distal, podem soar tanto óbvias quanto absurdas (atender

sem checar, checar cuidadosamente, fugir pelos fundos etc.). Casos como este parecem demandar que o próprio norte utilizado varie em função das circunstâncias. A relevância-t é, ela mesma, uma meta-dependência contextual. Isso faz com que o desafio maior à TR se apresente no mesmo espaço em que ele se apresenta nas demais teorias: como constituir e manter um contexto-t adequado, sendo que os mecanismos responsáveis por tal precisam apresentar sensibilidade ao contexto distal, e o uso de um norte fixo, como sugerem os autores, caracteriza a negação dessa sensibilidade e, consequentemente, da flexibilidade demandada?

O fato de que o mesmo padrão oriundo da adoção de alguma concepção de racionalidade limitada pode ser encontrado nestas três teorias não é suficiente, claro, para negar T5. É possível que ainda se formule alguma versão de uma tal teoria que supere este desafio. Contudo, o fato de que todas as teorias aqui analisadas apresentarem um problema de mesma natureza que constitui, cada qual ao seu modo, uma reformulação do FP, parece suficiente para justificar um pessimismo a esse respeito. Este é, ao mesmo tempo, um ponto a favor e um ponto contra Fodor. O ponto a favor é que esta conclusão parece reforçar sua predição de que toda tentativa de resolver o FP no âmbito da TRM resultaria em uma reformulação do problema. Por outro lado, o ponto contra é que ela enfraquece a conexão direta que ele estabelece entre o FP e a racionalidade, abrindo espaço para teses que caracterizem o problema em outros termos. Pode ser que, ao contrário do que sugeriu Fodor, o FP não esteja no nível da análise da da própria racionalidade. A natureza do problema, seja qual for, pode estar em outro lugar, talvez ainda mais profundo.

Conclusão

Como foi possível perceber, espera-se, no decorrer de toda a discussão, o ato de descrever o FP de um dado modo constitui, por si só, uma hipótese. Como tal, ela demanda esforço argumentativo, e é notório que, mesmo após todo o caminho percorrido, não há total clareza ainda sobre exatamente que tipo de problema é o FP. Tivesse sua natureza sido revelada, teria sido possível resolvê-lo ou, pelo menos, compreendê-lo de modo preciso o suficiente para que as condições sob as quais ele pode ser solucionado se mostrem salientes. Sequer isso foi realizado. Posto nestes termos, este parece o prenúncio de uma conclusão um tanto decepcionante. Contudo, disso não se segue que não tenha havido qualquer progresso ou que o processo não tenha sido instrutivo. Permitir que algo se revele como mais complicado do que inicialmente se supunha é, muitas vezes, um resultado satisfatório, ao menos na filosofia. Nesse sentido, e tal como previsto, o FP foi aqui encarado como tendo papel potencialmente semelhante ao que o ceticismo exerce na epistemologia: talvez ele possa ser vencido em definitivo, talvez não, mas isso não significa que ele não constitua uma ferramenta útil para jogar luz sobre aspectos inicialmente obscuros, ainda que importantes, em uma teoria da cognição que faça uso de representações.

Para explicitar o resultado obtido, convém relembrar e sintetizar o método utilizado para descrever o problema. Tomou-se como ponto de partida um amplo e vago problema geral da relevância (PGR). No interior deste, foram identificados dois elementos que, ao menos inicialmente, seriam característicos do FP: o requisito computacional e o requisito antropocêntrico. Tais elementos descrevem, ao mesmo tempo, as condições em que o problema se apresenta, e os critérios de especiação do PGR. Assim, dizer que o FP tem um requisito computacional é dizer que a teoria cognitiva em questão precisa ser computacional para que ele se mostre. Da mesma forma, dizer que ele tem um requisito antropocêntrico é dizer que o FP se mostra na tentativa de descrever o aparato cognitivo tal como ele é no ser humano, distinguindo-o, por exemplo, de algumas versões do FPIA. Ao fim do primeiro capítulo, apresentou-se aquilo que caracterizaria o principal instrumento de análise das teorias consideradas nos capítulos posteriores: o *requisito organizacional*, a ser somado aos outros dois já identificados. Dizer que o FP é caracterizado por esse requisito é dizer a solução deve envolver uma descrição sucinta do modo como a informação no interior do sistema é armazenada e organizada. Foi com base nesse requisito que se deu a distinção entre o aspecto inferencial e o aspecto organizacional do FP.

Na mesma linha, ao tratar dos efeitos contextuais possíveis no âmbito do veículo do pensamento, caracterizou-se uma conexão entre a ocorrência de meta-dependências contextuais e a ocorrência do FP na TRM. Essa conexão foi reiteradamente utilizada nas análises posteriores, a partir das noções de contexto proximal e distal. Com a

articulação desse ferramental, pôde-se sumarizar dois pontos fundamentais à investigação: primeiro, a enormidade do desafio que a TRM tem a vencer, condensada na noção de sensibilidade ao contexto distal, ou seja, na capacidade de lidar com meta-dependências contextuais. Segundo, a descrição das limitações das ferramentas que a TRM tem à disposição e o modo como tais limitações constituem seu principal obstáculo nessa empreitada, o que fora sintetizado no conceito de *eliminativismo contextual*.

Não apenas jogar, mas manter a luz sobre estes pontos de modo contínuo é importante porque, mesmo hoje, persiste uma conexão muito forte entre o trabalho que se realiza no âmbito da ciência cognitiva e as pesquisas desenvolvidas em IA. Contemporaneamente, a possibilidade de lidar com conjuntos de circunstâncias razoavelmente complexas num único sistema de IA (um sistema que dirige carros em vários ambientes, por exemplo) contribui para que a questão do tratamento do contexto pelas ciências cognitivas seja, em geral, tomada não como um problema fundamental, mas como uma dificuldade de implementação. Ainda que o desenvolvimento de tais sistemas não seja constrangido por um requisito antropocêntrico, persiste a noção vaga de que implementar o modo específico como o ser humano lida com o mundo demandaria "apenas" realizar uma versão muito mais complexa e computacionalmente demandante daquilo que já se produz em domínios específicos. Isso fortalece uma tendência segundo a qual as limitações e dificuldades são atribuídas ao ferramental tecnológico, e não ao ferramental conceitual. É importante reconhecer o quanto esta pode ser uma saída válida: uma técnica hoje considerada ineficaz pode vir a soar boa o suficiente, a depender do poder computacional disponível. Contudo, ela pode também ser fruto de uma compreensão insuficientemente profunda do que está em jogo. Para um exemplo, basta lembrar de Pollock: ao tratar da seleção sobre o que deve ou não ser considerado candidato à ponderação racional em tomadas de decisão, ele enxergou ali uma questão implementacional difícil para a qual não tinha uma resposta acabada, mas que poderia ser resolvida com o desenvolvimento de técnicas eficazes. Em contraste, essa mesma questão foi aqui apresentada não como um problema de implementação, mas sim como o problema mais fundamental e desafiador de sua teoria cognitiva, visto que tal seleção pressupõe a capacidade de resolver meta-dependências contextuais.

Além dessa conexão com a IA contemporânea, deve-se enfatizar que a resiliência de um obstáculo como o FP nem sempre se mostra de modo imediato e claro. Não é por acaso que foi necessário dar tanto espaço à caracterização do problema: uma compreensão rasa pode levar rapidamente a uma tentativa de solução que o subestime. Por isso foi realizado aqui um esforço argumentativo para mostrar a fundamental importância do FP, visto que a cognição humana é também sensível ao contexto distal. Este é o pano de fundo no qual toda atividade cognitiva humana se dá, e o modo como ele pode ser modelado permanece um mistério em função do seu caráter holístico. Ao tentar modelar o contexto distal como um conjunto de informações estruturadas, surge o FP, seja por exigir que o agente percorra computacionalmente, uma a uma, todas as potencialmente infinitas possibilidades, seja por apelo a estratégias *ad hoc* que ten-

tam circunscrever o processo de modo injustificável face ao requisito antropocêntrico. O problema específico da TRM, argumentou-se, é que ela está presa ao eliminativismo contextual: não parece haver outra saída exceto tratar do contexto distal como um conjunto de representações estruturadas. Embora ela consiga lidar com contextos quando estes são tomados como modelos formais delimitados a domínios específicos (contextos proximais), isso permite a ela tratar apenas de alguns efeitos contextuais (tais como indexicais), mas não todos. Vem daí a resiliência do FP, ou ao menos para este efeito se argumentou. Como se pôde notar no decorrer do desenvolvimento do trabalho, há vários exemplos de discussões que subestimam esse aspecto do problema, colaborando para que ele tenha deixado de ser, por um bom tempo, uma questão importante na filosofia da mente e na ciência cognitiva, mesmo entre os defensores da TRM.

O FP assim compreendido constitui uma métrica importante, tanto para quem quer defender, quanto para quem quer rejeitar a TRM. O primeiro tem no FP uma medida precisa da real capacidade que sua teoria cognitiva tem de lidar com todas as formas de sensibilidade ao contexto. Por essa via, ele pode tanto ampliar a compreensão da natureza do FP quanto melhor elucidar os requisitos que uma teoria da cognição adequada precisa atender. Este uso do problema como métrica é o que foi exemplificado no segundo e terceiro capítulos desse trabalho. Neles, analisou-se tanto a busca de uma solução pelo uso de gêneros representacionais distintos, quanto pela revisão do papel da racionalidade como guia na caracterização da inteligência humana. Em ambos os casos, foi possível notar problemas fundamentais com as estratégias adotadas. Estes problemas, argumenta-se, não seriam tão facilmente visíveis se a análise não tivesse sido realizada sob a luz da compreensão específica do FP aqui apresentada e de sua relação com a sensibilidade ao contexto. Desse modo, ao mesmo tempo em que isto permitiu uma descrição mais acurada do desafio que cada teoria tem a vencer, foi possível fornecer razões para pensar que o FP não é um problema caracterizado pelo uso de um certo gênero representacional (seja ele sentencial, pictorial ou distribuído) nem um problema fundamentalmente caracterizado pelo papel que a racionalidade exerce em teorias cognitivas. Esses casos constituem exemplos de como a compreensão do papel que o FP desempenha no arcabouço de uma teoria cognitiva permite melhor elucidar os pontos fortes e fracos destas teorias, ao menos diante do desafio dado pela sensibilidade ao contexto.

Por sua vez, também aquele que quer rejeitar a TRM, tem no FP uma ferramenta importante. Existem, claro, diversos caminhos para negar a TRM. O mais óbvio deles é desenvolver uma alternativa que não sofra do FP ao mesmo tempo em que mantenha consigo as virtudes da teoria rejeitada. Até onde foi possível investigar, contudo, uma tal opção está ainda por ser formulada. Há, portanto, bons motivos para querer fazer uso do FP como um argumento contra a TRM: ainda que não se tenha alternativa plausível a oferecer, é possível afirmar que a TRM está fadada ao fracasso, visto que ela não consegue superar o FP.

Contudo, ao contrário do que talvez possa parecer em princípio, o FP não constitui um argumento direto pelo abandono do uso de representações na explicação do conteúdo intencional de um agente. Um tal argumento teria como pressuposto uma elucidação completa da natureza do problema. É possível, como se viu, descartar algumas possibilidades (o FP não parece ser um problema essencialmente ligado ao uso de um certo gênero representacional, por exemplo), mas a ideia de que o FP é fruto direto do uso de representações, tenham elas a natureza que tiverem, carece de justificativa, pois seria preciso apontar, de modo preciso, qual característica da representação mental é responsável pelo surgimento do problema.

A literatura contemporânea fornece vasto material para formular várias hipóteses desse tipo (ou seja, esse é um trabalho que poderia ter inúmeros capítulos adicionais). A natureza da relação entre representações mentais e processos computacionais é um exemplo pertinente: quanto do que se toma como características fundamentais de uma representação mental não poderia ser fruto de um pressuposto oculto acerca do como elas são utilizadas em processos com caráter computacional? Veja-se a separação, onipresente em todas as teorias aqui abarcadas, entre o mecanismo que armazena uma representação e o mecanismo que faz uso dela para algum fim. Essa é uma demanda computacional, não representacional. Uma tal distinção poderia, em princípio, constituir um critério de especiação adicional (para além do gênero) de diferentes tipos de representações. Pode a natureza do FP estar ali oculta? Este pode ser, sem dúvida, objeto de investigações futuras.

Nesse cenário, aquilo que se sabe acerca do FP interessa tanto ao que quer defender, quanto ao que quer rejeitar a TRM. A ambos importa a questão: para onde ir a partir deste ponto? A conclusão a que esse trabalho conduziu permite eliminar alguns caminhos que seriam pouco frutíferos, mas não permite gerar, por si mesma, nenhuma proposta positiva de hipótese a perseguir. Estas devem chegar a partir da conjunção do que fora aqui desenvolvido com discussões que se dão em outros espaços no interior da filosofia e da ciência cognitiva. O desafio está em detectar quais, dentre esses elementos externos, são relevantes.

Bibliografia

ANDLER, D. From Paleo- to Neo-connectionism. In: VIJVER, Gertrudes Van De (Ed.). **New Perspectives on Cybernetics: Self-Organization, Autonomy and Connectionism**. Dordrecht: Springer Science & Business Media, 1992. v. 220p. 125–146.

ANDLER, D. Is Context a Problem? **Proceedings of the Aristotelian Society**, v. 93, n. 1, 01 jun. 1993.

ANDLER, D. The normativity of context. **Philosophical Studies**, v. 100, n. 3, p. 273–303, 2000a.

ANDLER, D. Context and background. Dreyfus and cognitive science. In: WRATHALL, Mark; PALMAS, Jeff (Eds.). **Heidegger, Coping, and Cognitive Science: Essays in Honor of Hubert L. Dreyfus**. Cambridge: The MIT Press, 2000b. p. 137–159.

ANDLER, D. Context: the case for a principled epistemic particularism. **Journal of Pragmatics**, v. 35, n. 3, p. 349–371, 2003.

BECHTEL, W. **Mental Mechanisms**. New York: Routledge, 2008.

BIANCHI, C. Three forms of contextual dependence. In: **International and Interdisciplinary Conference on Modeling and Using Context**. [s.l.] Springer, 1999. p. 67–76.

CAMP, E. Thinking with Maps. **Philosophical Perspectives**, v. 21, n. 1, 2007.

CHALMERS, D.; CLARK, A. The Extended Mind. **Analysis**, v. 58, n. 1, 1998.

CHEMERO, A. **Radical embodied cognitive science**. Cambridge, MA: MIT Press, 2009.

CHERNIAK, C. **Minimal Rationality**. [s.l.] MIT Press, 1990.

CHIAPPE, J., D. L.; Vervaeke. Fodor, Cherniak and the Naturalization of Rationality. **Theory & Psychology**, v. 7, n. 6, 01 dez. 1997.

CHOW, S. J. What's the Problem with the Frame Problem? **Review of Philosophy and Psychology**, v. 4, n. 2, p. 309–331, 2013.

CHURCHLAND, P. M. **A Neurocomputational Perspective: The Nature of Mind and the Structure of Science**. Cambridge: The MIT Press, 1989.

CLAPIN, H. **Philosophy of mental representation**. [s.l.] Clarendon Press, 2002.

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

Press, 1998.

CLARK, C., Andy; Thornton. Trading spaces: Computation, representation, and the limits of uninformed learning. **Behavioral and Brain Sciences**, v. 20, n. 01, mar. 1997.

CROCKETT, L. J. **The Turing Test and the Frame Problem in AI**. [s.l.] Intellect Ltd, 1994.

CUMMINS, R. **Meaning and mental representation**. Cambridge: MIT Press, 1991.

CUMMINS, R. Inexplicit Information. In: **The World in the Head**. [s.l.] Oxford, 2010. p. 86–97.

DENG, L.; YU, D. **Deep Learning: Methods and Applications**. [s.l.] NOW Publishers, 2014.

DENNETT, D. Cognitive wheels: the frame problem of AI. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 41–64.

DENNETT, D. Inteligência artificial como filosofia e como psicologia. In: **Brainstorms: ensaios filosóficos sobre a mente e a psicologia**. Tradução Luiz Henrique De Araújo Dutra. [s.l.] Unesp, 1999[1978]. p. 163–183.

DENNETT, D. C. **The intentional stance**. [s.l.] MIT press, 1989.

DIETRICH, C., Eric. And Fields. The role of the frame problem in Fodor's modularity thesis: a case study of rationalist cognitive science. **Journal of Experimental & Theoretical Artificial Intelligence**, v. 7, n. 3, p. 279–289, 1995.

DREYFUS, H. **What computers can't do: a critique of artificial reason**. New York: Harper & Row, 1972.

DREYFUS, H. L. Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. In: **Artificial Intelligence**. [s.l.] Elsevier, 2007. v. 171p. 1137–1160.

DREYFUS, H. L.; DREYFUS, S. E. How to stop worrying about the frame problem even though it's computationally insoluble. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 95–111.

DUPRE, J. **Processes of Life: Essays in the Philosophy of Biology**. 1. ed. New

York: Oxford University Press, 2012.

FODOR, J. A. **The Language of Thought**. [s.l.] Harvard University Press, 1980.

FODOR, J. A. **Representations**. Sussex: The Harvester Press, 1981. v. 13

FODOR, J. A. **The Modularity of Mind**. [s.l.] MIT Press, 1983.

FODOR, J. A. Modules, frames, fridgeons, sleeping dogs and the music of the spheres. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 139–149.

FODOR, J. A. **Concepts: Where Cognitive Science Went Wrong (Oxford Cognitive Science Series)**. [s.l.] Oxford University Press, 1998.

FODOR, J. A. **The mind doesn't work that way: the scope and limits of computational psychology**. [s.l.] A Bradford Book, 2001.

FODOR, J. A. **LOT 2: The Language of Thought Revisited**. [s.l.] Oxford University Press, 2010.

FODOR, J. A.; PYLYSHYN, Z. W. Connectionism and cognitive architecture: A critical analysis. **Cognition**, v. 28, n. 1-2, p. 3–71, 1988.

GELDER, T. Van. Defining "Distributed Representation". **Connection Science**, v. 4, n. 3-4, jan. 1992.

GELDER, T. Van. Dynamics and cognition. In: HAUGELAND, John (Ed.). **Mind design II: phylosophy, psychology, artificial intelligence**. [s.l.] MIT Press, 1997. p. 421–450.

GLYMOUR, C. Android epistemology and the frame problem: comments on Dennett's "Cognitive Wheels". In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 65–75.

GRAVES, A.; WAYNE, G.; DANIHELKA, I. Neural turing machines. **arXiv preprint arXiv:1410.5401**, 2014.

GRICE, P. **Studies in the Way of Words**. [s.l.] Harvard University Press, 1991.

HADLEY, R. F. Systematicity in Connectionist Language Learning. **Mind and Language**, v. 9, n. 3, p. 247–272, 1994.

HASELAGER, W.; RAPPARD, J. V. Connectionism, Systematicity, and the Frame Problem. **Minds and Machines**, v. 8, p. 161–179, 1998.

HAUGELAND, J. An overview of the frame problem. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex,

1987. p. 76–93.

HAUGELAND, J. **Artificial Intelligence: The Very Idea**. [s.l.] Bradford, 1989.

HAUGELAND, J. Mind embodied and embedded. In: HAUGELAND, John (Ed.). **Having thought**. Cambridge: Harvard University Press, 1998a. p. 207–237.

HAUGELAND, J. The nature and plausibility of cognitivism. In: HAUGELAND, John (Ed.). **Having thought**. Cambridge: Harvard University Press, 1998b. p. 9–45.

HAUGELAND, J. Representational Genera. In: **Having thought**. Cambridge: Harvard University Press, 1998c. p. 171–206.

HAYES, P. The frame problem and related problems in artificial intelligence. In: ELITHORN, A.; JONES, D. (Eds.). **Artificial and human thinking**. [s.l.] Jossey-Bass, 1973. p. 41–59.

HAYES, P. J. **In defence of logic**. Proceedings IJCAI 77. **Anais**. . . 1977

HAYES, P. J. What the frame problem is and isn't. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 123–137.

HEBB, D. O. **The Organization of Behavior: A Neuropsychological Theory**. [s.l.] Taylor; Francis, 2002 [1949].

HENDRICKS, S. The frame problem and theories of belief. **Philosophical studies**, v. 129, n. 2, p. 317–333, 2006.

HOLYOAK, K. J.; HUMMEL, J. E. The proper treatment of symbols in a connectionist architecture. In: DIETRICH, Erik; MARKMAN, Art (Eds.). **Cognitive dynamics: Conceptual change in humans and machines**. [s.l.] MIT Press, 2000. p. 229–263.

HUME, D. **Tratado da natureza humana**. Tradução Déborah Danowski. São Paulo: Unesp, 2000.

HUMMEL, J. E.; HOLYOAK, K. J. Distributing structure over time. **Behavioral and Brain Sciences**, v. 16, n. 03, set. 1993.

HUMMEL, K. J., John E.; Holyoak. Distributed representations of structure: A theory of analogical access and mapping. **Psychological Review**, v. 104, n. 3, 1997.

JANLERT, L.-E. Modeling Change - The Frame Problem. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 1–40.

JANLERT, L.-E. The frame problem: freedom or stability? With pictures we can have both. In: FORD, Kenneth M.; PYLYSHYN, Zenon W. (Eds.). **The robot's dillemma**

revisited: The frame problem in artificial intelligence. [s.l.] Ablex, 1996. p. 35–48.

JOHNSON-LAIRD, P. N. Mental models in cognitive science. **Cognitive science**, v. 4, n. 1, p. 71–115, 1980.

JOHNSON-LAIRD, P. N. Mental models and human reasoning. **Proceedings of the National Academy of Sciences**, v. 107, n. 43, 26 out. 2010.

MARCUS, G. F. Rethinking Eliminative Connectionism. **Cognitive Psychology**, v. 37, n. 3, 1998.

MCCARTHY, J. Circumscription - a form of non-monotonic reasoning. **Artificial intelligence**, v. 13, n. 1, p. 27–39, 1980.

MCCARTHY, J.; HAYES, P. J. Some philosophical problems from the standpoint of artificial intelligence. **Machine Intelligence**, v. 4, p. 463–502, 1969.

MCCLELLAND, J. L. et al. **Parallel distributed processing**. Cambridge, MA: MIT press, 1987. v. 2

MCDERMOTT, D. We've been framed: or, why AI is innocent of the frame problem. In: PYLYSHYN, Zenon W. (Ed.). **The robot's dillemma: The frame problem in artificial intelligence**. [s.l.] Ablex, 1987. p. 113–122.

MCDERMOTT, D.; DOYLE, J. Non-monotonic logic I. **Artificial intelligence**, v. 13, n. 1-2, p. 41–72, 1980.

MCGINN, C. **The character of mind**. [s.l.] Oxford University Press Oxford, 1982.

MINSKY, M. A framework for representing knowledge. In: HAUGELAND, John (Ed.). **Mind design II: phylosophy, psychology, artificial intelligence**. [s.l.] MIT Press, 1997. p. 111–142.

NEWELL, A.; SHAW, J. C.; SIMON, H. A. **Report on a general problem solving program**. IFIP congress. **Anais**. . . Pittsburgh, PA, 1959

NILSSON, R. E. F. N. J. Strips: A new approach to the application of theorem proving to problem solving. **Artificial Intelligence**, v. 2, n. 3-4, 1971.

PALMER, S. Fundamental aspects of cognitive representation. In: ROSCH, Eleanor; LOYD, B. B. (Eds.). **Cognition and categorization**. Hillsdale NJ: Erlbaum, 1978.

PERINI-SANTOS, E. Contextualismo. In: BRANQUINHO, João; SANTOS, Ricardo (Eds.). **Compêndio em linha de problemas de filosofia analítica**. [s.l.] Centro de filosofia da universidade de Lisboa, 2014.

PICCININI, G. Some neural networks compute, others don't. **Neural Networks**, v. 21, n. 2-3, 2008.

POLLOCK, J. L. **Thinking about acting: logical foundations for rational decision**

making. [s.l.] Oxford University Press, USA, 2006.

RADFORD, A.; METZ, L.; CHINTALA, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. **CoRR**, v. abs/1511.06434, 2015.

RECANATI, F. **Perspectival Thought: A Plea for Moderate Relativism**. [s.l.] Oxford University Press, 2007.

REITER, R. A logic for default reasoning. **Artificial intelligence**, v. 13, n. 1-2, p. 81–132, 1980.

RESCORLA, M. Cognitive Maps and the Language of Thought. **The British Journal for the Philosophy of Science**, v. 60, n. 2, 01 jun. 2009.

RESCORLA, M. From Ockham to Turing–and Back Again. In: **Philosophical Explorations of the Legacy of Alan Turing**. Cham, Switzerland: Springer, 2017. p. 279–304.

ROBINSON, J. A. A machine-oriented logic based on the resolution principle. **Journal of the ACM (JACM)**, v. 12, n. 1, p. 23–41, 1965.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. **Psychological Review**, p. 65–386, 1958.

RUMELHART, D. E. et al. **Parallel distributed processing**. Cambridge, MA: MIT press, 1987. v. 1

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

SALAY, N. **Why Dreyfus' Frame Problem Argument Cannot Justify Anti-Representational AI**. Proceedings of the Annual Meeting of the Cognitive Science Society. **Anais**. . . 2009

SAMUELS, R. Classical computationalism and the many problems of cognitive relevance. **Studies in History and Philosophy of Science**, v. 41, n. 3, p. 280–293, 2010.

SANDEWALL, E. An approach to the frame problem and its implementation. **Machine intelligence**, v. 7, n. 195-204, p. 11–19, 1972.

SCHANK, R. C. **The primitive ACTs of conceptual dependency**. Proceedings of the 1975 workshop on Theoretical issues in natural language processing. **Anais**. . . Association for Computational Linguistics, 1975

SEARLE, J. R. Literal meaning. **Erkenntnis**, v. 13, n. 1, p. 207–224, 1978.

SHANAHAN, M. **Solving the Frame Problem: A Mathematical Investigation of the Common Sense Law of Inertia**. [s.l.] MIT, 1997.

SHANAHAN, M. The Frame Problem. In: ZALTA, Edward N. (Ed.). **The Stanford Encyclopedia of Philosophy**. Spring 2016 ed. [s.l.] https://plato.stanford.edu/archives/spr2016/

entries/frame-problem/; Metaphysics Research Lab, Stanford University, 2016.

SHASTRI, L.; AJJANAGADDE, V. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. **Behavioral and Brain Sciences**, n. 16, p. 417–494, 1993.

SIMON, H.; NEWELL, A. Heuristic problem solving: the next advance in operations research. **Operations Research**, v. 5, n. 1, p. 1–10, 1958.

SMOLENSKY, P. Tensor product variable binding and the representation of symbolic structures in connectionist systems. **Artificial Intelligence**, v. 46, n. 1-2, 1990.

SPERBER, D. Modularity and Relevance: How Can a Massively Modular Mind Be Flexible and Context-Sensitive. In: CARRUTHERS, Peter; LAURENCE, Stephen; STICH, Stephen (Eds.). **The Innate Mind: Structure and Contents**. [s.l.] Oxford University Press, 2005. p. 53–68.

SPERBER, D.; WILSON, D. **Relevance: communication and cognition**. [s.l.] John Wiley; Sons, 1995.

SPERBER, D.; WILSON, D. Fodor's frame problem and relevance theory. **Behavioral and brain sciences**, v. 19, n. 3, p. 530–532, 1996.

STERELNY, K. **The Representational Theory of Mind**. Oxford: Basil Blackwell, 1990.

THELEN, E.; SMITH, L. **A dynamic systems approach to the development of cognition and action**. Cambridge, MA: MIT Press, 1992.

TIENSON, J.; HORGAN, T.; TIENSON, J. (EDS.). **Connectionism and the Philosophy of Mind**. Dordrecht: Kluwer Academic Publishers, 1991.

TURING, A. M. On computable numbers, with an application to the Entscheidungsproblem. **Proceedings of the London mathematical society**, v. 2, n. 1, p. 230–265, 1936.

TVERSKY, A.; KAHNEMAN, D. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. **Psychological review**, v. 90, n. 4, p. 293, 1983.

VICENTE, A. Context dependency in Thought. In: RECANATI, Francois; STOJANOVIC, Isidora; VILLANUEVA, Neftali (Eds.). **Context-dependence, perspective and relativity**. [s.l.] De Gruyter, 2010. p. 68–89.

WASKAN, J. A. Intrinsic cognitive models. **Cognitive Science (Elsevier Science)**, v. 27, n. 2, 2003.

WASKAN, J. A. **Models and Cognition: Prediction and Explanation in Everyday Life and in Science**. Cambridge: MIT Press, 2006.

WHEELER, M. **Reconstructing the Cognitive World: The Next Step (MIT Press)**.

[s.l.] A Bradford Book, 2007.

WHEELER, M. Cognition in context: phenomenology, situated robotics and the frame problem. **International journal of philosophical studies**, v. 16, n. 3, p. 323–349, 2008.

WINOGRAD, T. Understanding natural language. **Cognitive Psychology**, v. 3, n. 1, 1972.

WINSTON, P. **Artificial Intelligence**. [s.l.] Addison-Wesley, 1992.

ZIEMKE, T. What's that thing called embodiment. In: ALTERMAN; KIRSH (Eds.). **Proceedings of the 25th Annual Conference of the Cognitive Science Society**. Mahwah, NJ: Lawrence Erlbaum, 2003. p. 1305–1310.