FAKULTÄT FÜR !NFORMATIK

Renewing Cognitive Science

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

im Rahmen des Studiums

Computational Intelligence

eingereicht von

Friedrich Slivovsky
Matrikelnummer 0202583

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer: Ao.Prof.Dipl.-Ing.Dr.techn. Christian Georg Fermüller

Wien, 02.04.2009

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Friedrich Slivovsky
Sobieskigasse 18/5
1090 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 02. 04. 2009 _____

(Unterschrift Verfasser/in)

Kurzfassung

Die Kognitionswissenschaft erhebt den Anspruch, eine einheitliche Theorie des Denkens zu formulieren, in der sich die Anstrengungen solch unterschiedlicher Disziplinen wie AI, Anthropologie, Philosophie, Psychologie und der Neurowissenschaften bündeln. Insbesondere geht man davon aus, die philosophische Frage nach dem *Wesen* von Intelligenz durch naturwissenschaftliche Untersuchung und Modellierung der *Funktionsweise* intelligenter Systeme klären zu können. Dieser methodische Zugang ist jedoch überaus fragwürdig, da philosophische Begriffsklärung und empirische Wissenschaft *prima facie* als voneinander unabhängig zu sehen sind. Ich beabsichtige zu zeigen, dass das dominante kognitionswissenschaftliche Paradigma, in dessen Zentrum eine Computer-Metapher steht (*Computational-Representational Understanding of Mind*), für diese Identifikation naturwissenschaftlicher und philosophischer Fragestellungen verantwortlich ist. Darüber hinaus werde ich die beiden diesem Ansatze zugrundeliegenden Konzepte (*Computation* bzw. *Representation*) einer kritischen Analyse unterziehen, um einige gravierende begriffliche Probleme aufzuzeigen. Wie sich herausstellen wird, ist die Kognitionswissenschaft allgemein von *funktionalistischen* Motiven durchdrungen. Ein Blick auf die Funktionalismuskritik Hilary Putnams wird zeigen, dass diese philosophische Theorie unhaltbar ist.

Auf der Suche nach möglichen Alternativen werde ich mich Robert Brandoms *Inferentialismus* zuwenden, und eine knappe Zusammenfassung seiner Begriffstheorie vorlegen. Ich werde die These aufstellen, dass sein *normativer Pragmatismus* dazu beitragen könnte, einige innerhalb der Kognitionswissenschaft herrschende Missverständnisse zu beseitigen. Neben der Beantwortung philosophischer Fragen könnte Brandoms Projekt auch einem veränderten Verständnis naturwissenschaftlicher Forschung im Zusammenhang mit Kognition Vorschub leisten, das die Bedingungen für die Teilnahme an der sozialen Praxis des Gebens und Verlangens von Gründen in den Vordergrund stellt.

TECHNISCHE
UNIVERSITÄT
WIEN

VIENNA
UNIVERSITY OF
TECHNOLOGY

Renewing Cognitive Science

by

Friedrich Slivovsky

Submitted to the Faculty of Informatics in Partial
Fulfillment of the Requirements for the Degree of

Master of Science

in

Computational Intelligence

at the
Vienna University of Technology

Supervisor: Ao.Prof.Dipl.-Ing.Dr.techn. Christian Georg Fermüller

Vienna, April 2, 2009

Abstract

Cognitive science purports to offer a unified theory of the mind, combining research from such disparate fields as AI, anthropology, philosophy, psychology, and neuroscience. In particular, the philosophical question of *what* is distinctive of intelligence is supposed to be answered by developing models of *how* cognitive systems produce intelligent behavior. I will argue that this methodological commitment is fallacious, because philosophical explication and empirical research constitute independent areas of inquiry. Identifying the dominant research paradigm within the discipline – referred to as the *Computational-Representational Understanding of Mind* (CRUM) – as the source of this confusion, I will take a closer look at both the notions of *representation* and *computation*, pointing out several conceptual problems. Contending that standard theories in cognitive science are variations of a *functionalist* theme, I will appeal to arguments by Hilary Putnam, providing evidence for the inadequacy of functionalism as a philosophical theory of cognition.
In an attempt to present an alternative, I will turn to Robert Brandom's *inferentialism*. I will offer glimpses of this compelling conception of the nature of cognition, and assess the prospects of a cognitive science inspired by it. Following Brandom, I will argue that a *normative pragmatist* foundation for cognitive science could resolve the methodological problems inherent to the discipline. By answering to the philosophical question of *what* cognition is, it further gives rise to a novel understanding of the empirical question, asking *how* systems manage to produce the kind of social linguistic behavior that counts as "giving and asking for reasons."

Contents

1 Introduction **5**
- 1.1 Cognitive Science . . . 6
- 1.2 Representations, Computation, and Functionalism . . . 9
- 1.3 AI: An Engineer's Perspective . . . 11

2 Computation **17**
- 2.1 Cummins and Schwarz . . . 18
- 2.2 Chalmers . . . 22
- 2.3 Churchland and Sejnowski . . . 24
- 2.4 Reverse Engineering . . . 28
- 2.5 Cognitive Functions . . . 32
- 2.6 Summary . . . 33

3 Representation **34**
- 3.1 The Classical Case . . . 35
- 3.2 Connectionism . . . 36
  - 3.2.1 Implementationalism . . . 37
  - 3.2.2 Eliminativism . . . 38
  - 3.2.3 The Case for Constituent Structure . . . 39
- 3.3 Summary . . . 44

4 Cognitive Science and Functionalism **45**
- 4.1 "Mindfulness" . . . 46
- 4.2 Inference and Intelligence . . . 48
- 4.3 Functionalism . . . 52
  - 4.3.1 Computation and Representation . . . 53
- 4.4 What Functionalism can't do . . . 56
  - 4.4.1 Semantic Externalism . . . 56
  - 4.4.2 "Narrow" Content and "Broad" Content . . . 56
  - 4.4.3 A Theory of Reference . . . 57
- 4.5 Summary . . . 58

5 Cognitive Science and Inferentialism **59**
- 5.1 Introduction . . . 59
  - 5.1.1 Sapience . . . 60
- 5.2 Normative Pragmatism . . . 61

5.3 Semantic Inferentialism 62
5.4 Deontic Scorekeeping 64
5.4.1 The Nature of Norms 65
5.4.2 Norms of Deontic Scorekeeping 67
5.5 Normative Attitudes and Normative Status 68
5.6 Norms and Supervenience 71
5.7 A New Perspective for Cognitive Science 73

Bibliography **77**

1 Introduction

From its inception, artificial intelligence (AI), broadly understood as a discipline concerned with building machines displaying some form of *intelligence*, was accompanied by a program which sought to advance a psychological and philosophical theory of cognition. Or rather, these two tasks were thought of as two sides of the same coin: if it were possible to understand the mind in terms of a *computer metaphor*, machine implementations that could reproduce intelligent performances with respect to a specific domain would yield promising candidates for a psychological theory, and – vice versa – a successful (computational) theory of cognition could – in principle – be implemented in silicon.

For a while, AI made good progress, and experts in the field regularly estimated that they were getting closer and closer to the creation of machines exhibiting intelligence as encountered in human beings (referred to as *general intelligence*). In time, however, the research program faced serious obstacles. The early successes achieved in detached micro-worlds could not be extended to more realistic scenarios, and the exhaustive formalization of the background knowledge and holistic practices which appear to guide our understanding seemed infeasible. Ultimately, it became apparent that hopes had been too optimistic, predictions too bold, and the project of classical AI started to disintegrate.

The concomitant philosophical project known as *(machine) functionalism* – an attempt to construe mental states as *functional* states within an algorithm – became subject of substantial criticism as well (see Putnam, 1988), although not as an immediate consequence of the fate of AI. As usual, philosophical theories die hard, and functionalism is still very much alive. Attempts to directly arrive at a program which would confer on its algorithmic states the semantic contentfulness required for intelligence may have failed so far, but in the eyes of functionalists that in itself does not rule out the *possibility* that such programs exist.

The collapse of classical symbolic AI gained further momentum from the emergence of *connectionism* as an alternative paradigm, offering neural network models which did particularly well at a range of tasks where good-old-fashioned (symbolic) AI (GOFAI) had failed miserably, such as pattern recognition. But the story of the demise of classical AI has been told elsewhere (see Dreyfus, 1979) and is not the focal point of this work. For the present purposes, the history of AI is relevant only in tracing the roots of the discipline with which we will mainly be concerned: the field known as *cognitive science*.

With ad-hoc approaches failing to provide the desired results, theorists of AI turned to what was envisaged as "no-tricks basic study" (Dreyfus, 1979, p.27). In its first issue, the *Cognitive Science Journal* described the agenda of the new field

as follows:

> Cognitive science is defined principally by the set of problems it addresses and the set of tools it uses. The most immediate problem areas are representation of knowledge, language understanding, image understanding, question answering, inference, learning, problem solving, and planning. ... The tools of cognitive science consist of a set of analysis techniques and a set of theoretical formalisms. The analysis techniques include such things as protocol analysis, discourse analysis, and a variety of experimental techniques developed by cognitive psychologists in recent years. The theoretical formalisms include such notions as means-end analysis, discrimination nets, semantic nets, goal-oriented languages, production systems, ATN grammars, frames, etc.
> **(Collins, 1977)**

Along with the theoretical underpinnings of AI, cognitive science was conceived as harboring the philosophical and psychological projects that originated in its vicinity. The field is sometimes falsely identified with its psychological aspects, viz. the question of *how* humans think, but it is equally concerned with the fundamental conceptual issues of *what* it means for something to think or be intelligent. These two issues are fused in a methodological approach that seeks to **explain what cognition is by exploring models of how it is done**.

It is this particular methodological commitment that constitutes the main target of the criticism worked out in this thesis: I will argue that these problems must be addressed individually, and offer an explanation of why cognitive science blurs the line separating the respective domains.

But before going into the specifics of this endeavor, a few general remarks are in order. Even though the present work is to be regarded an investigation into the conceptual foundations of cognitive science, its intended audience includes those more closely affiliated with parts of AI that are not concerned with the philosophical issues debated here. Therefore, part of this introduction is devoted to an attempt at convincing this group that philosophical insights into the nature of cognition may yield dividends for the theory of creating "intelligent" artifacts. Therefore, following a brief introduction into cognitive science and an outline of the issues discussed in subsequent chapters, I will (again, briefly) indicate how philosophy may inspire AI. In addition to that, I will try to defend a stronger thesis, according to which a conception of AI as completely isolated from philosophical questions (what I want to call an "engineering" approach) is inconsistent with some of the discipline's goals.

1.1 Cognitive Science

To motivate a cognitive scientific understanding of the mind, let me identify two important aspects of intelligence, or, to put it more carefully, two properties nor-

mally attributed to intelligent beings, both related to what is called *intentionality*. With the failure of classical behaviorist research programs, it has become widely accepted within psychology and the philosophy of mind that intelligent beings *really* have mental states, intentions, beliefs, etc. Usually, we can associate contents with these cognitive states, in the sense that it is legitimate to ask what they are *about* (what is *represented* by them). *That* mental states are subject to semantic interpretations of this kind, I believe, is a fact on which most theorists agree, and competing paradigms only have different ways of explaining it.

But simply *having* these cognitive states is merely one aspect of intentionality as encountered in intelligent beings. We cannot imagine someone going through different states of this kind in a completely arbitrary manner – clearly, there are systematic relations between mental states based on their contents. For instance, my being in an intentional state that involves being thirsty and perceiving a glass of water will normally lead to forming an intention of drinking that glass of water, where "normally" just means something like "in absence of reasons not to" (e.g. believing that the water is contaminated).

Moreover, what an individual believes in, thinks, or feels, should somehow affect their actions – in other words, intentional states ought to be causally significant to behavior. In case of the water glass, the intention of drinking should result in appropriate actions. Again, we can envision scenarios where someone does not act in accordance with his or her state of mind. How exactly to conceive of these systematic relations is a delicate matter that need not concern us at this point. What is important is that outside such a systematic context, intentional states would be rendered entirely unintelligible.

Now, I am not claiming that a combination of these features I have mentioned gives us a definition of intelligence, but any appropriate theory of intelligence certainly needs to address them in one way or another. That is, it has to offer an account of

(a) what it is for an individual to be in an intentional state

(b) how these intentional states form systematic relations

Cognitive science, broadly speaking, offers answers to both of these questions in terms of *representations* and *computation*. Accordingly, to be in an intentional state is to have a certain mental representation or to stand in a specific relation to it. Further, computational procedures operating on these representations are invoked in explaining the systematicity governing the succession of intentional states and their logical interconnection. Essentially, this is the characterization put forward by Paul Thagard in *Mind: Introduction to Cognitive Science* (Thagard, 2005):

> Here is the central hypothesis of cognitive science: Thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures. Although there is much disagreement about the nature of the representations

> and computations that constitute thinking, the central hypothesis is general enough to encompass the current range of thinking in cognitive science, including connectionist theories. For short, I call the approach to understanding the mind based on this centrals hypothesis *CRUM*, for *Computational-Representational Understanding of Mind*.
> **(Thagard, 2005, p.10)**

This amounts to a generalized computer-metaphor: *representations* are modeled on data structures, while *computations* correspond to algorithmic procedures. AI, for the most part, may have abandoned this paradigm for understanding the mind, but it is still considered a live option – even state of the art – within cognitive science. I must hasten to add that, at least according to cognitive scientists themselves, the discipline's interpretation of this approach is far more liberal than the one associated with classical AI. In this context, theorists emphasize both the variety of specific mind-models subsumed by cognitive science and the interdisciplinary character of the field, presumably uniting efforts from such disparate areas as AI, anthropology, philosophy, psychology, and neuroscience. The interdisciplinary take on the subject is introduced as one of the major virtues of cognitive science, but it highlights a major philosophical problem. It is indisputable that all of the aforementioned disciplines in some way contribute to our understanding of mental phenomena. But cognitive science's aspiration to a "unified view", a "common interpretation of how the mind works" (Thagard, 2005, p.19) clearly goes far beyond this trivial observation. What is envisaged is nothing less than *one* set of concepts, *one* source of explanatory primitives, presumably provided by CRUM. But exactly how is this goal thought to be achieved? Psychology, in accounting for human performances, avails itself of an *intentional* idiom, including such notions as "belief" and "desire." Neuroscientific theories, on the other hand, are couched in a language describing (for instance) biochemical processes surrounding neural activity. These vocabularies belong to entirely different domains, and an attempt of reducing them to a common paradigm runs in danger of merely changing the subject.

I intend to focus on the relation between two other areas of inquiry, represented by two questions, that cognitive science seeks to offer a "unified" answer for. These are, on the one hand, the philosophical (conceptual) question of *what* cognition amounts to, of what kinds of entities we talk about using intentional language, and the (broadly) empirical question of *how* cognitive "systems" produce intelligent behavior, of how they realize the required cognitive capacities, on the other.

I will argue that these issues need to be held apart – going even further, one must answer to the first question to be in a position to answer the second. This idea is taken from a paper by Robert Brandom (Brandom, 2008b), whose philosophical project will be introduced in chapter 5, as offering an alternative to the current conceptual underpinnings of mainstream cognitive science. It may appear as straightforward that unless one has an account of *what* cognition is, attempts of explaining *how* the relevant properties are attained by cognitive systems are without determinate direction. But, as I mentioned above, the approach pursued by

cognitive science can informally be characterized as "explaining what cognition is by exploring models of how it is done."

I will now turn to a brief analysis of the appeal of this methodological approach, and explain how cognitive science's aspiration to a unified theory of cognition should be understood. Subsequently, I will provide an outline of the contents of this thesis, and set the agenda for the remaining chapters.

1.2 Representations, Computation, and Functionalism

The remark I made earlier about cognitive science conceiving of intentional states as mediated by *representations* remained fairly abstract. To make this idea more substantial, think of the intimate relation between linguistic expressions and cognition: thoughts have contents that can be expressed as claims, *representing* states of affairs. In this way, we may conceive of a claim expressing the content of an intentional state as a kind of *representation*. Therefore, if we identify intentional states with claims articulating their contents, the succession of intentional states can be thought of as a sequence of *representations*. As contended above, the mind passes through such representational states in a systematic, reason-respecting manner. These are *philosophical* statements, and, as far as I see it, fairly uncontroversial.

Cognitive science, in claiming that these representations are (physically) encoded in the mind/brain, and positing algorithmic *computations* to account for the systematicity governing their succession, first and foremost puts forward a fascinating *psychological* theory. How, then, is it supposed to answer to philosophical questions about the mind at the same time, as part of a unified theory of cognition?

The internal representations appealed to in cognitive scientific theory are subject to semantic interpretation, but *prima facie* they can be conceived as purely syntactic items. One of the animating ideas of classical AI was that, rather than being conferred on the system from outside, the semantics is generated by the syntactic relations themselves, in the spirit of Haugeland's well-known motto: "if you take care of the syntax, the semantics will take care of itself" (Haugeland, 1985, p.106). In other words, the meanings attributed to tokens are taken as implicit in the rules governing their manipulation. Transposed to a somewhat more general setting, the project of construing semantic contents in terms of functional roles is known as *functionalism*. Applying this view to intentionality, the contents attributed to intentional states result from their being caught up in suitable computational relations. That is to say, for something to be a *belief* is to assume a state in an algorithm, systematically linked to other algorithmic states corresponding to beliefs.

Here, the computational model figures not just in an account of *how* intelligent behavior is possible, but similarly of *what* it means for behavior to be intelligent. It is particularly attractive as it purports to offer a thoroughly materialistic under-

standing of the mind: mental states are computational states, and the mind/brain *implements* the corresponding algorithm.

Accordingly, cognitive systems are identified with computational systems, or a particular subclass of computational systems. It is not just that this thesis in itself is controversial – in some sense, it is not even clear what the controversy is about: standard explications of **computation** suffer from profound conceptual problems. In general, the understanding of the concept is too vague, resulting in a proliferation of "computers." More restrictive interpretations, I will argue, conflict with the discipline's self-understanding as discovering novel forms of computation, while liberal interpretations run in danger of rendering the notion vacuous. It will turn out that the resulting friction is intimately related to questionable aspects of the field's methodology.

Let me illustrate how the functionalist coupling of internal mechanisms and intentional states generates a new set of philosophical problems. Assume we contrived a mind-model incorporating as representations ordinary linguistic expressions describing beliefs, intentions, etc., in combination with a computational routine performing logical operations on them, "inferring" further beliefs (or nonlinguistic behavior). One problem with this sort of approach is that one often notices what appear as "gaps" in conscious reasoning – it is difficult to reconstruct the underlying inferential steps. The data available through introspection is partial at best, canceling out the possibility of directly deriving the envisaged algorithm.

Rather than refining the reasoning-routine, we may be inclined to respond by modifying our set of representations, moving to presumably more fine-grained "subconscious" psychological states. The desired representations are then located somewhere between conscious reasoning and the nexus of electrical signals constituting brain function at the bottom level. That is not to say there *really* is a continuum – but instead of positing representations whose implementation is problematic, why not directly examine physiological structures in an attempt to reveal their presumed semantics? At any rate, once the identity of representations and ordinary sentences describing mental states is gone, things get complicated. If we can best explain intelligent behavior without appeal to ordinary mental states, should psychology dispense of the latter altogether?

Taking a somewhat different approach to the same problem, one could ask: if these representations do not correspond to ordinary linguistic expressions, what is the relation between these two species of content-bearers? Is the standard arsenal of psychological concepts we employ just an approximation to theories located on a subconscious level? Should cognitive theory abandon ordinary language in favor of this more exact idiom? From within a functionalist frame of mind, there seem to be only two options: either reconstruct ordinary intentional vocabulary in terms of internal representations, or abandon this mode of explanation (often referred to as "folk psychology").

For that reason, where **representations** are invoked, the contents attributed to them are *conceptual* contents, putatively approximating and accounting for the contents of intentional states. These contents are at the same time conceived as

resulting from the functional roles assigned to individual representations within the cognitive model, giving rise to a thoroughly functionalist understanding of intentional states.

The chapter on **functionalism** will be devoted mainly to a condensed summary of Hilary Putnam's critical assessment of this philosophical project in *Representation and Reality* (Putnam, 1988). I will further raise the question as to why functionalism, in spite of its deficiencies, is still considered a viable position within the philosophy of cognitive science, and suggest that this status derives from a perceived lack of alternatives: it is assumed that functionalism offers the only account of the mind compatible with a scientific world-view.

In particular, using Hilary Putnam's phrase, there appears to be a "horror of the normative" (Putnam, 2004, p.70). By contrast, in the final chapter, I will turn to Robert Brandom's theory of cognition (more precisely: of concept use) that centers on the insight that intentional states have a primarily *normative* significance, determining what those to whom they are ascribed *ought* to do. Because these proprieties of behavior can be interpreted as underwriting *inferences*, his project is known as **Inferentialism**.

Following Brandom, I will argue that a normative foundation for cognitive science could resolve the methodological problems inherent to the discipline. By answering to the philosophical question of *what* cognition is, it further gives rise to a novel, pragmatist understanding of the empirical question, asking *how* systems produce the kind of social linguistic behavior that counts as "giving and asking for reasons." I will offer glimpses of this compelling conception of the nature of cognition, and assess the prospects of a cognitive science inspired by it.

1.3 AI: An Engineer's Perspective

Some of these results directly carry over into artificial intelligence: they contribute to establishing standards intelligent artifacts have to satisfy, and help identify the capacities required to meet them. If the explanatory target of cognitive science becomes more explicit, so do design goals in AI.

To provide further evidence for the relevance of philosophical insights to AI, I will now describe a theoretical position I will refer to as the **engineering perspective**. Although I am not entirely convinced that it represents a position actually assumed by those working in the discipline, I think it is the only philosophically coherent perspective to take if one intends to avoid philosophical discussions altogether. Yet, as I will argue, it comes at the price of imposing severe restrictions on the scope of AI.

I will proceed by characterizing a set of important problems associated with AI that *cannot* be articulated from within this frame of mind. Whether losing the ability to specify these problems is to be regarded an impoverishment of the discipline, ultimately depends, of course, on whether they are thought of as essential to its scientific program.

Aspirations of revealing the principles governing intelligence aside, building and evaluating computational models is an integral part of any account of AI. For a large part, work in the discipline consists in exploring specific methods and their application to formally regimented problem domains. Accordingly, AI appears as a motley of tools and models rather than an integrated scientific paradigm devoted to unraveling the mystery of cognition. If research within these distinct areas can proceed autonomously, wouldn't it be possible to rid AI of questions regarding the philosophical (or psychological) relevance of its results and focus exclusively on its technical aspects? After all, the theory of neural networks can be framed without the burden of connectionism, and courses on mathematical logic usually do not involve an introduction to epistemology. Such an ascetic conception of the discipline definitely has its virtues, avoiding the philosophical pitfalls that come with interpreting AI as providing theories of intelligence. To put it in a (slightly polemical) slogan: "Let cognitive science worry about the psychological and philosophical implications of our models, while we engage in substantial research and solve real world problems!"

As tempting as it seems, I believe completely ignoring philosophical discussions on matters of cognition would restrain AI more than most people from the field wish for. At a certain point, the (legitimate) aim to dispose of problematic philosophical commitments that came with GOFAI collides with the discipline's interest in areas such as natural language processing and computer vision.

To see why this is the case, the attitude reflected in statements like the one above must first be expressed in way that allows for analysis. That puts me in a somewhat uncomfortable position: to view through philosophical lenses a position which is distinctly *a*philosophical may seem a little absurd. But it is more of an inconvenience than a flat-out self-contradiction. I am not re-importing a philosophical dimension only to criticize it later on. Instead, I am trying to extract from conceptions of AI which reject philosophical theorizing a coherent point of view and then explore its implications. In consequence, the force of this argument is limited to the degree that the explication successfully captures what is implicit in such accounts of AI. I believe, however, that *if* one wants to avoid philosophical discussions regarding intelligence altogether, the following is the only conceivable route to take.

In order to characterize this *engineering perspective* on AI, let me introduce the notion of an (intelligent) *agent*. According to Russell's and Norvig's standard textbook on AI titled *Artificial Intelligence: A Modern Approach*,

> an **agent** is anything that can be viewed as **perceiving** its environment through **sensors** and **acting** upon that environment through **effectors**.
> **(Norvig and Russell, 1995, p.31)**

Many accounts of AI turn on this concept, defining the purpose of the discipline as the design and study of agents – more specifically, *intelligent* or *rational* agents. The crux of this definition is filling in an appropriate idea of intelligence or rationality:

> A **rational agent** is one that does the right thing. Obviously, this is better than doing the wrong thing, but what does it mean? As a first approximation, we will say that the right action is the one that will cause the agent to be most successful. That leaves us with the problem of deciding *how* and *when* to evaluate the agent's success.
> **(Norvig and Russell, 1995, p.31)**

Here is a way of avoiding the notorious debates and arrive at an *engineering perspective*: although AI revolves around *intelligent* agents, we can substitute for this notion of intelligence (or rationality) criteria of an agent's *success* in solving the particular task at hand.

> Therefore, we will insist on an objective performance measure imposed by some authority. In other words, we as outside observers establish a standard of what it means to be successful in an environment and use it to measure the performance of agents.
> As an example, consider the case of an agent that is supposed to vacuum a dirty floor. A plausible performance measure would factor in the amount of electricity consumed and the amount of noise generated as well. A third performance measure might give highest marks to an agent that not only cleans the floor quietly and efficiently, but also finds time to go windsurfing at the weekend.
> **(Norvig and Russell, 1995, p.31)**

The *intelligence* of agents, then, consists in their living up to some *objective performance measure* defined relative to a limited domain. The bottom line is a retail, rather than wholesale conception of intelligence: instead of requiring a uniform standard of rationality, what counts as intelligent is decided in an *ad hoc* fashion, possibly involving experts in the particular field.

Now, to put it in more rigid terms: *the engineering perspective on AI places at the heart of the discipline the design of agents whose behavior conforms to an objective performance description. The latter must be rendered in a naturalistic language (eschewing, in particular, cognitive or intentional concepts).*

The restriction imposed on the language used to characterize the performance measure is meant to rule out situations in which an account of general intelligence returns in a detailed description of the agent's behavior. For example, building an agent whose success is identified with a high number of "reasonable" answers to a set of questions is outside the scope of the engineering perspective as defined here. (It may be argued that Russell and Norvig are trying to make a similar point by insisting on an "objective" performance measure.)

Without any such constraint, the AI-engineer might find herself under obligation to offer interpretations of concepts such as "intelligent" in the course of designing an agent, and, as far as I understand it, that is precisely what the attitude underlying the engineering perspective seeks to avoid.

Against the background of this characterization, let me try to explain what difficulties this specific approach entails. First of all, it is unlikely that an objective performance measure is available to the AI-theorist independently from the design process. That is to say, it may not always be possible to pass on the responsibility of arriving at a suitable set of criteria to an expert acquainted with the problem domain.

Most of the time, the engineer will be included in the processes creating the agent's specification. And as part of the modeling process, one applies concepts of "intelligence" or "intelligent behavior", inevitably shaping the final result. Accordingly, the goal of these design activities, the formal requirements on the agent's or system's behavior, will often amount to a characterization of intelligence with respect to the problem domain. That is to say, although the final result may be couched in "naturalistic" vocabulary, the performances which one thereby seeks to capture are precisely those that would count as *intelligent* in the original context.

Treating the specification as if it were conceptually opaque to AI is in danger of merely relocating the problems that can arise from incomplete or fallacious accounts of rationality. Translating the latter to viable objective performance measures may not be trivial even for scenarios that are rather clear-cut, as the following example by Russell and Norvig illustrates:

> There is a danger here for those who establish performance measures: you often get what you ask for. That is, if you measure success by the amount of dirt cleaned up, then some clever agent is bound to bring in a load of dirt each morning, quickly clean it up, and get a good performance score.
> **(Norvig and Russell, 1995, p.31)**

Delegating the task of formalizing intelligent behavior in the context of specific applications is a legitimate move, but it does not necessarily improve the quality of the resultant agents.

But the engineering perspective's main flaw consists in its limited scope. For a large range of problems, the aforementioned restrictions concerning the language in which to specify the agent's behavior are without consequences. As long as any occurrence of controversial vocabulary can ultimately be translated to an objective performance measure, the corresponding task falls in the domain of the engineering perspective.

But there are problems – problems which have traditionally been of great importance to AI – that have so far resisted attempts of description in purely naturalistic terms. These are sometimes referred to as *AI-complete* problems (Wikipedia, 2009) and include natural language processing (automated translation, information extraction) and computer vision. They are thought to represent the hardest class of problems in AI, solving which requires so called *strong AI* – intelligence at the level of human cognition.

Let me suggest an alternative way of identifying AI-complete problems: any (reasonably detailed) description of what an agent designed to solve them needs

to do appeals to notions of *intelligence* that cannot be translated[1] to an objective performance measure in the above sense.

I do not want to pretend I have a knock-down argument supporting this claim, but there is conclusive evidence:

- Consider what is known as the *holistic* character of intelligence: mastering one aspect of reasoning usually involves mastering others. As a consequence, it is impossible to decompose our cognitive capacities into a set of individually intelligible faculties, which could in turn be described in a language devoid of intentional concepts.

- If there *was* a description of any AI-complete problem in naturalistic vocabulary, solving it would amount to a – more or less – straightforward engineering task. (That is to say, if it can be solved at all. Following a similar line of thought, the failure of arriving at strong AI has raised doubts in some theorists as to the computability of the "functions" in question. I believe this is a dead end: we need a better understanding of the capacities that are supposed to be tackled computationally, rather than exploring novel paradigms of computation.)
  But so far, these problems have resisted every attempt of solving them.

If this is correct, having a philosophical theory of *intelligence* (or a more narrowly construed faculty associated with it) is *necessary* to even describe this range of tasks within AI. Specifically, what I have called the *engineering perspective* lacks the conceptual resources to express what is required of systems designed to solve these problems.

As I said before: it is a legitimate position for AI to avoid the sticky philosophical issues related to synthesizing intelligence, but to take it precludes one from confronting the hardest problems traditionally associated with artificial intelligence. Should those affiliated with AI care? After all, most of them were not in pursuit of strong AI to begin with. Lowering the inflated expectations that came with GOFAI, far from being disastrous, might allow the discipline to focus on realistic problems instead of chasing windmills. But as things stand, coming to grips with the day-to-day problems AI faces is likely to be inseparable from studying general intelligence – yet another consequence of the holistic character of cognition. There appears to be no continuum of intelligence, in the sense that one could start by studying or creating moderately intelligent systems and iteratively increase their complexity as to arrive at general intelligence. When it comes to intelligence, in *some* sense, it seems to be all-or-nothing.

[1] A word of caution is advised here: chapter 5 will be devoted to articulating (partially, at least) what an agent has to do in order to count as deploying concepts and engaging in discursive practice. As a part of this, certain notions are explicated by using a vocabulary which is assumed more basic for the given purpose (what Brandom refers to as *pragmatic metavocabulary* (Brandom, 2008a)). But having a description of what it is an agent has to do in order to count as intelligent in this sense does not mean we can insert that description for every occurrence of the word "intelligent" in the specification of an AI-complete problem.

That is not to suggest that interpretations of AI with an explicit philosophical agenda (see Poole et al., 1998) and those expressed by what I have called the engineering perspective represent the only options. If the arguments brought forward in this thesis are correct, the former are no viable choice. Yet, as I shall argue in the concluding chapter, we need not therefore confine ourselves within the narrow boundaries drawn by the latter. The *via media* is given by an approach which renounces the ambitions to advance a computational theory of cognition, but retains the objective of synthesizing general intelligence, by paying attention to a number of important lessons drawn from the philosophy of mind.

2 Computation

The concepts of *representation* and *computation* are closely linked. If the fact that individuals proceed through intentional states in a reason-governed manner is to be explained by invoking *computation*, then *representations* must already be in place. Just as algorithmic procedures rely on the presence of data structures whose content they modify, *computation* depends on *representation*. Representationalism (roughly: the idea that intentional states derive their semantics from content-bearing tokens in the mind) does not entail computationalism, but computationalism certainly involves representationalism – there is "no computation without representation" (Thagard, 2005, p.153).

In this section, I will study attempts by cognitive scientists to get a grip on the fundamental concept of *computation*. Despite its pivotal role within the theoretical framework of the discipline, there seems to be no standard account of its meaning that is agreed upon.

> It is perhaps worth mentioning that the foundational notion of computation is itself still surprisingly ill understood. What do we really mean by calling some phenomenon "computational" in the first place? There is no current consensus at least (in the cognitive scientific community) concerning the answer to this question. It is mostly a case of "we know one when we see one."
> **(Clark, 2001, p.17)**

Just as with *representation*, the notion of *computation* is usually conceived to be very general by cognitive scientists. According to such interpretations, it serves as a metaphor encompassing both symbolic algorithms proper and various other ways of manipulating representations, for instance, the kind of systematic changes to weight-vectors employed in connection with neural networks. Just as patterns of activation in such networks can be viewed as a kind of *representation*, their rule-governed modification is regarded a species of *computation*.

On such a reading, connectionism is just another, if unusual, instance of computationalism. Considering the effort that connectionists have often put into setting their paradigm apart from the classical picture, it is startling that both theories should fit into a common conceptual framework, one whose outlines are clearly shaped by symbolic approaches. It is even more surprising, however, that this interpretation need not be read into connectionism but is one that theorists in the field adhere to themselves. Faced with challenges brought forward by proponents of the symbolic approach, connectionists have dedicated much work to reconcile their

views with the conceptual framework established by classical AI and cognitive science. They themselves have made attempts of defining notions of *representation* and *computation* applicable to neural networks.

This view is not universally shared, and some theorists are concerned that it is distortive of the connectionist agenda, possibly undermining its attempts to open up novel perspectives on understanding the mind. The worry is that this way, rather than leaving behind classicists views, connectionism may be usurped by the philosophical idiosyncrasies inherent to the symbolic paradigm.

Be that as it may, the "computationalist" interpretation of connectionism appears to be the dominant theoretical perspective within the field. Incidentally, two of the texts on which the subsequent discussion of *computation* is based are written by authors affiliated with this camp. Although this fact is, as we shall see shortly, virtually irrelevant given the notion's consistency across the classical/connectionist divide, it is worth explicit mentioning to avoid certain critical responses. It may be objected that the critique worked out below merely provides evidence for the inadequacy of connectionism, at least in its computationalist incarnation.

Against such worries, I must stress the fact that it is not my intention here to argue in favor of or against connectionism, or, for that matter, specifically computationalist interpretations of connectionism (even though, of course, the critical assessment of computationalism may render such interpretations less viable). Regardless of whether (computationalist) connectionism is true, its characterization of fundamental notions of computationalism seem to be accurate (and representative of classical theories), and those are the primary focus of this critique.

Studying connectionist accounts of *computation* further yields a substantial methodological payoff: many issues concerning the *computational-representational understanding of mind* are even more visible against the background of computationalist connectionism than they are in context with the classical symbolic paradigm. Once the computer-metaphor becomes metaphorical itself (because neural networks are blatantly not computers in an ordinary sense), its predicaments are revealed ever so clearly. Where classicists could operate with concepts like *representation* and *computation* (more or less) in a straightforward manner, connectionists first have to reflect on the meaning of these notions to see whether they are applicable to their models. Consequently, the assumptions underlying computationalist approaches, often implicit in the classical case, are brought into the open for assessment.

2.1 Cummins and Schwarz

In a paper titled "Connectionism, Computation, and Cognition" (Cummins and Schwarz, 1991), Robert Cummins and Georg Schwarz engage in the endeavor of locating mainstream connectionism within the broader field of what they call "computationalism" – the hypothesis that "systems are cognitive in virtue of computing

appropriate functions" (60). [1]

In order to illustrate their concept of computation (or, in this case, "calculation"), they appeal to the following example:

> During an archaeological expedition you unearth a strange looking device. Close examination reveals that the device has a lid which hides twelve buttons, each with a strange symbol etched on its face, and a small window. Curiosity leads you to experiment, and you notice that certain sequences of button pushings result in certain symbols being displayed in the window, displays which consist only of symbols that also appear on the buttons. ... The display uses only ten of the twelve little symbols found on the buttons, so you start mapping them onto the ten digits familiar from our decimal numeral system. ... The rest is easy, for a few test runs show that system's output (its display state) can be consistently interpreted as the product of the numbers that interpret the antecedent sequence of button pushings. As it turns out, your ancient device is a multiplier.
> **(Cummins and Schwarz, 1991, pp.60-61)**

Indeed, under these circumstances, most of us would probably be inclined to call the device a calculator (or, more specifically, a multiplier). But why is it that we interpret the alien artifact as a calculator, rather than something else? Schwarz and Cummins seem to be clear about the fact that there are no easily discernible intrinsic (physical) properties which warrant this interpretation. The answer they propose involves what the call an "interpretation function" – a mapping between physical states (61) and representational states, establishing a relation between representations of numbers (and – if we examine the interior of the device – state transitions) and what physically corresponds to them. As for the current example, what makes the artifact a multiplier is the possibility to identify physical (display) states with the arguments and result of the product function, respectively. Additionally, those states – revealed as representational states in virtue of an appropriate interpretation function – ought to be causally significant:

> A multiplier is a device such that causing it to represent a pair of numbers causes it to represent their product. Thus, to explain how a device multiplies is (at least) to explain how representations of multiplier and multiplicand cause representations of (correct) products.
> **(Cummins and Schwarz, 1991, p.61)**

In order to identify some of the conceptual problems inherent to this definition, we need to examine a crucial distinction introduced by the authors, according to

[1] It is worth noting that the authors do indicate the possibility of a "non-computationalist connectionism" – turning on their definition of computationalism, however, rather than leaving it behind. Moreover, even this characterization of connectionism is firmly rooted within representationalism.

which *computing* a function needs to be set apart from merely *satisfying* it. For instance, "calculators satisfy the MULTIPLY(n1,n2) function by computing it. A falling apple, on the other hand, satisfies the function $D = (at^2)/2$ but does not compute it. What's the difference?" (62) They quickly come up with an answer, involving execution of an algorithm for the function f in question:

> The natural, familiar and, we believe, correct suggestion about computing is that computing a function **f** is executing an algorithm that gives **o** as its output on input **i** just in case **f(i)=o**. The problem of explaining function computation reduces to the [problem] of explaining what it is to execute an algorithm for that function. The obvious strategy is to exploit the idea that algorithm execution involves steps, and to treat each elementary step as a function that the executing system (or one of its components) simply satisfies. ... So: to compute a function is to execute an algorithm, and algorithm execution is disciplined step satisfaction.
> **(Cummins and Schwarz, 1991, p.62)**

That is, a system or device *satisfies* a function **f** in case certain parts (or states) interpretable – by means of an interpretation function – as its input **i** and output **o**, respectively, are systematically related by the equation **f(i)=o**. It *computes* **f** if, further, upon examining its internal features, we are able to identify components *satisfying* functions corresponding to steps in an algorithm for **f**. Thus, at the bottom level, *function computation* collapses to *function satisfaction*.

What is wrong with that definition? First of all, it doesn't quite tell us how to pick out distinctively *computational* systems, simply because, in its generic form, it is too loose. In order to get any grip on this class of systems, it needs to be augmented with some reasonably precise definition of the term *algorithm*. Otherwise, what operations is the cognitive theorist allowed to identify as algorithmic "steps" the system *satisfies*? Of course, some suitable definition could be put forward. For instance, one could identify *algorithms* with *Turing Machines*. But the brain evidently is not *literally* a Turing Machine, and whatever operations we can locate in its physiology do not directly correspond to moves of a read/write head on an infinite tape (it seems highly implausible that cognitive processes can be perspicaciously interpreted as Turing Machines on *any* level of description). On a further note, bear in mind that results establishing the equivalence of formal computer models provide no comfort here – there is no reason to assume that if some artifact computes the function **f** by going through states corresponding to steps in one formalism thereby also goes through states corresponding to the "simulation" of the steps computing **f** in another formalism.

But let us assume that a reasonable definition could be given – possibly as the disjunction of individual formalisms – that puts us in a position to decide, for any particular state transition in the physical system, whether it qualifies as a possible step in an algorithm. Assume we intend to apply this refined definition of *computation* to an arbitrary device we do not yet know is a computer. Should we

regard it a computer if there is *some* conceivable algorithm it implements? Even on a very restrictive definition of *algorithm* (that is in danger of *excluding* some computational systems), there *will* be *some* such algorithm for *any* physical system whose structure we choose to study. As a consequence, according to this definition, *everything* would have to be identified as a *computer*. But clearly, that is not the intention of Cummins and Schwarz:

> There are lots of causal processes, and only some of them are instances of function computation. It is the latter that constitute the realm of computational explanation. Moreover, there are many ways of satisfying functions, even computable functions (i.e., functions for which there are algorithms) that do not involve executing an algorithm.
> **(Cummins and Schwarz, 1991, p.63)**

However, I think it is clear that their account of the notion is hardly capable of maintaining this demarcation.

This line of criticism has been followed by several prominent authors in the philosophy of mind. In an appendix to his review of functionalism, *Representation and Reality* (Putnam, 1988), Hilary Putnam presented an argument effectively proving that *every finite state automaton is implemented by every physical system.*[2] John Searle strikes a similar note in an essay titled "Is the brain a digital computer?" (Searle, 1990), in context with what he calls "universal realizability":

> On the standard textbook definition of computation,
>
> (a) For any object there is some description of that object such that under that description the object is a digital computer.
>
> (b) For any program there is some sufficiently complex object such that there is some description of the object under which it is implementing the program.
>
> **(Searle, 1990)**

The bottom line of these criticisms is that *computation*, according to the prototypical definitions encountered in cognitive science, is not an "objective" property, or, even worse, a completely vacuous notion. If *every* physical system has the property of being a *computational* system, the identification of such systems has no explanatory value.

We may attempt to fix this issue by imposing further restrictions, thus getting rid of the vagueness inherent to the isomorphism-based model. Taking a closer look at its structure, we can identify three integral elements:

(a) **states or parts of a physical system**

[2] More precisely, his theorem states that "every ordinary open system is a realization of every abstract finite automaton." (Putnam, 1988, p.121)

(b) **states of an algorithm**

(c) **an isomorphism relating 1 and 2**

Arguably, the isomorphism is implicit in one's choice of 1 and 2. But while it is straightforward to conceive of an algorithm in terms of interdependent functional states, it is not trivial to carve up the physical system in a similar manner. I take it that one would have to **(a)** properly choose time-slices that correspond to individual states, in addition to **(b)** picking out parts of the system relevant in encoding functional states. Apparently, it is tacitly assumed that it is obvious how to make these choices, but why? I believe the rationale is that, because the envisaged mapping of computational states to physical states exists in the case of ordinary computers, something similar must be true in general. Regular computational systems are isomorphic to algorithms in virtue of their being *designed* to implement these particular algorithms and encode the corresponding algorithmic states. But unfortunately, we cannot conclude from the existence of this one *characteristic* isomorphism that it is the *only* conceivable mapping. On the contrary, the arguments put forward by Searle and Putnam boil down to there being *too many* possible isomorphisms.

The most natural way to face this problem is to restrict the space of possible choices by regimenting one's interpretation of "physical state" and "algorithmic state" in a reasonable manner.

2.2 Chalmers

A rather sophisticated version of this strategy was presented by David Chalmers in *The Conscious Mind* (Chalmers, 1996). He maintains that, in spite of the problems concerning universal realizability mentioned above, "an objective account of implementation[3] can be given" (316). Instead of Turing Machines, Chalmers chooses so called *Combinatorial State Automata* (*CSA*s) as formal algorithmic model, but this is a minor technical detail. Otherwise, his exposition of the notion of *implementation* sounds rather familiar:

> Informally, we say that a physical system *implements* a computation when the causal structure of the system mirrors the formal structure of the computation. That is, the system implements the computation if there is a way of mapping states of the system onto states of the computation so that physical states that are causally related map onto formal states that are correspondingly formally related.
> **(Chalmers, 1996, pp.317-318)**

[3] In Chalmer's terminology, "implementation" roughly corresponds to the notion of *computation* as it is used here, while "computation" refers to an abstract, mathematical level.

The *formal* definition offered further in the text adds nothing new to this, only rephrasing this description in mathematical terms for any specific CSA M and physical system P. Formally, this exposition is virtually identical to the one discussed above. In contrast to Cummins's and Schwarz's account, however, he introduces further restrictions in breaking down the physical system under examination:

> We may stipulate that in a decomposition of the state of a physical system into a vector of substates, the value of each element of the vector must supervene on a separate region of the physical system, to ensure that the causal organization relates distinct components of the system. Otherwise, it is not clear that the detailed causal structure is really present within the physical system.
> **(Chalmers, 1996, p.318)**

Clearly, there is still room for interpretation. But the constraints can be tightened as to allow for reasonably exact standards deciding whether an algorithm is implemented by a particular physical system. Put that way, it may be possible to immunize the definition of *computation* ("implementation") against the threat of universal realizability:

> What is crucial is that there is no reason to believe that *every* CSA will be implemented by *every* system. For any given CSA, very few physical systems will have the causal organization required to implement it.
> **(Chalmers, 1996, p.319)**

At any rate, the remaining degree of observer-relativity has certain undesirable consequences, as Chalmers is aware:

> It is true that *some* computations will be implemented by every system. For example, the single-element, single-state CSA will be implemented by every system, and a two-state CSA will be implemented almost as widely. It is also true that most systems will implement more than one computation, depending on how we carve up that system's states. There is nothing surprising about this: it is only to be expected that my workstation implements a number of computations, as does my brain.
> **(Chalmers, 1996, p.319)**

(The reference to *workstations* is obviously misleading. Although implementations of different algorithms may run on the same computer – even in parallel – that is not to say that it is a matter of "how we carve up that system's states", in the sense that it is up to *interpretation* what computation a certain machine realizes at any given moment. The capability to run "a number of computations" is the result of the machine's implementing *one specific* computation, rather than the consequence of varying interpretations.) Again, the consequence is a kind of *pancomputationalism*: every physical system implements *some* algorithm. Note that, apart from the trivial CSAs supposedly implemented by every system, there is at least another

one corresponding to a fine-grained description of its causal structure. Hence, even though Chalmers is able to preserve the notion of *computation* for purposes of characterizing a system's organization, his account runs counter to the commonsense intuition that *only some* things are *computers* (for that reason, it is also unable to maintain the kind of distinction envisaged by Cummins and Schwarz). Further, according to this story, the notion of *implementation* still has an interest-relative quality, albeit less significant. Searle contends that, on the standard definitions of *computation*, it is impossible to get rid of this surplus of interpretation:

> I think it is probably possible to block the result of universal realizability by tightening up our definition of computation. ... But these further restrictions on the definition of computation are no help in the present discussion because the really deep problem is that syntax is essentially an observer relative notion. The multiple realizability of computationally equivalent processes in different physical media was not just a sign that the processes were abstract, but that they were not intrinsic to the system at all. They depended on an interpretation from outside.
> **(Searle, 1990)**

One way of dealing with this problem is to confront it squarely, taking up the story about interpretation in an account of the concept. Identifying computational systems may require interpretation, but is that reason for concern? Why not embrace the freedom resulting from an inflation of concepts like *computation* and *computer*? On such views, rather than constituting a problem, observer-relativity is just part of the semantics of these notions.

2.3 Churchland and Sejnowski

This is the route taken by Churchland and Sejnowski in their explication of the term *computation* as part of their book titled *The Computational Brain* (Churchland and Sejnowski, 1992). In a fashion similar to Cummins's and Schwarz's, they intend to work out a definition of "computer" that is general enough as to apply to the brain (or neural networks in general) just as well as to ordinary computers. Although what they aim for is clearly a precise philosophical account of the concept *computation*, at the beginning of their remarks they modestly concede that

> the definition of computation is no more *given* to us than were the definitions of light, temperature, or force field. While some rough-hewn things can, of course, be said, and usefully said, at this stage, precision and completeness cannot be expected. And that is essentially because there is a lot we do not yet know about computation. Notice in particular that once we understand more about what sort of computers nervous systems are, and how they do whatever they do, we shall have an enlarged and deeper understanding of what it is to compute and

> represent.
> **(Churchland and Sejnowski, 1992, p.61)**

These lines articulate an objection to one of the central ideas motivating the present text, viz. that conceptual questions about cognition need to be answered prior to empirical research in cognitive science. In essence, Churchland and Sejnowski claim that this cannot be done, and their argument deserves closer examination.

As far as physics is concerned, their point is certainly valid. It would be naive to ask What is light? and exclude the best scientific theories available as possible sources for answers. The natural laws governing the propagation of light etc. are part of the conceptual content of the term "light" – one cannot conceive of conceptual and empirical questions in this regard as completely detached. Yet I am not convinced the situation perfectly carries over to "computation" in context with computational neuroscience (or cognitive science as a whole). Unlike light and force fields, computations, on Churchland's and Sejnowski's own account, are not strictly natural phenomena: a system instantiates a computation just in case it can be interpreted *as* computational. If we do not have an understanding of what "computation" means, what is the point of adopting this particular mode of analysis?

In contrast, at any given moment in the history of modern physics, physicists had a rather clear understanding of the concepts which were part of their theories (not taking into consideration competing paradigms). They may not have known as much *about* atoms, for instance, as we do today, but they would not have regarded such empirical knowledge a prerequisite to identifying atoms within their theories. No scientist would have said: "we don't have a clue what atoms are, but research will tell us."

What is particularly irritating about Churchland and Sejonwski's strategy is their appeal to the preliminary character of their knowledge in *making substantial claims*. The presumed vagueness concerning the notion of computation is perceived as legitimizing computational interpretations of the brain. To them, the issue is not *whether the brain is a computer at all*, but only *what sort of computer*. They even seem to be aware of the circularity this approach suffers from when they talk about the discipline as "bootstrapping itself up" (Churchland and Sejnowski, 1992, p.61). But the circle is vicious, rather than virtuous.

I will return to this issue after a more detailed presentation of their account of *computation*.

> Second, in the most general sense, we can consider a physical system as a computational system when its physical states can be seen as representing states of some other systems, where transitions between its states can be explained as operations on the representations.
> **(Churchland and Sejnowski, 1992, p.62)**

Further below, an attempt of a more formal definition is offered:

> let us hypothesize that a physical system computes some function **f** when (1) there is a systematic mapping from states of the system onto the arguments and values of **f**, and (2) the sequence of intermediate states executes an algorithm for the function.
> **(Churchland and Sejnowski, 1992, p.65)**

No surprises here. What sets their conception of *computation* apart from the ones discussed earlier is the status they assign to interpretation:

> We count something as a computer because, and only when, its inputs and outputs can usefully and systematically be interpreted as representing the ordered pairs of some function that interests us. Thus there are two components to this criterion: (1) the objective matter of what function(s) describe the behaviour of the system, and (2) the subjective and practical matter of whether we care what the function is.
> **(Churchland and Sejnowski, 1992, p.65)**

The second criterion is apparently supposed to somehow restrict application of the concept to relevant situations. But, if our "interest" decides what system can be regarded a computer, if this "subjective matter" is relevant, then, again, there are hardly any restrictions to what the concept subsumes.

If, on their conception of the concept, what counts as a computer is subject to "social or idiosyncratic conventions" (Churchland and Sejnowski, 1992, p.65) that hardly qualifies as a reconstruction of the meaning of this notion in ordinary contexts. Normally, it is simply not *up to our interest* whether something is a computer. Churchland and Sejnowski, however, explicitly endorse this aspect of their definition:

> It may be suggested as a criticism of this very general characterization of computation that it is *too* general. For in this very wide sense, even a sieve or a threshing machine could be considered a computer, since they sort their inputs into types, and if one wanted to spend the time at it, one could discover a function that describes the input-output behaviour. While this observation is correct, it is not so much a criticism as an apt appreciation of the breadth of the notion. It is rather like a lawn-growing perfectionist incredulously pointing out that on our understanding of "weed", even dandelions might be nonweeds relative to some clime and some tribe of growers. And so, indeed, they might be some farmer's cash crop.
> **(Churchland and Sejnowski, 1992, p.66)**

Clearly, the term "computer" normally isn't conceived general enough as to include threshing machines and sieves. It may be objected that it was not the authors' intention to explicate the use of the word in ordinary situations, but to introduce a notion that, although it certainly exhibits features of the former, is entirely internal to cognitive science. Consequently, it is of no interest that they do not offer a

precise account of the english word "computer." But in that case, there's no reason for them to argue in favor of the appropriateness of their concept. The notion they present is quite obviously parasitic on the colloquial usage of the word, and the whole project of conceiving of the brain in computational terms implicitly relies on such conventional understanding which is at the same time disqualified.

Let us refine the scenario surrounding the archaeological discovery of Cummin's and Schwarz's multiplier in the following way: imagine that a device physically identical to our multiplier had come into existence by a cosmic coincidence. Obviously, we can still discover an isomorphism between its physical states and the states in an algorithm for multiplication. But does that make it a multiplication device? Considering its merely being the result of an awkward contingency, is it reasonable to interpret it in these terms? It may not be straightforward to come up with an answer, but there is one implicit in the definitions of *computation* discussed so far: they clearly introduce objects into the realm of calculation devices which haven't been engineered with that purpose in mind.

> Not too much emphasis should be placed on the link between purposeful design and use as a computer, however, for a fortuitously shaped rock can be used as a sundial. This is a truly simple computer-trouvé, but we do have reason to care about the temporal states that its shadow-casting states can be interpreted as representing.
> **(Churchland and Sejnowski, 1992, p.66)**

What about objects which aren't (or never have been) used as computers? Apparently, the authors are willing to apply the term even in such remote cases.

> the system of Aubrey holes at Stonehenge computers eclipses of the sun by dint of the fact that its physical organization and state transitions are set up so that the sun stone, moon stone, and nodal stone land in the same hole exactly when an eclipse of the sun occurs. Notice that this would be so even in the highly unlikely event that Stonehenge was the fortuitous product of landslides and flooding rather than human contrivance.
> **(Churchland and Sejnowski, 1992, pp.66-67)**

Even if their definition is not meant as an elucidation of the ordinary term "computer", if its introduction results in such a proliferation of computers, it runs in danger of losing any utility the colloquial term might otherwise have and becoming idle vocabulary. After all, if *everything* is a computer, what, specifically, can we expect of examining the brain as if it were a computer? Churchland and Sejnowski seem to be aware of this problem, and, surprisingly, it is another interest-relative property that is supposed to confine application to reasonable situations:

> Finding a device sufficiently interesting to warrant the description "computer" probably also entails that its input-output function is rather

> complex[4] and inobvious, so that discovering the function reveals something important and perhaps unexpected about the real nature of the device and how it works.
> **(Churchland and Sejnowski, 1992, p.66)**

For the sake of the argument, let us grant Churchland and Sejnowski their conception of "computer" as interest-relative and up to interpretation. In some sense, they are not that far off: in the process of implementing an algorithm, physical states or properties are assigned meanings by the designer, and that *is* done in an ultimately arbitrary manner, even if severely constrained by technical factors. There still remains a deep ambiguity, manifest in the methodological remarks I quoted earlier. Computational neuroscience is supposed to enrich our understanding of the term "computer" by examining the brain in computational terms, but at the same time, it is the partial character of our grasp of the notion "computer" that is invoked as what permits the computational mode of analysis. There are, I believe, two valid strategies of applying a computational scheme:

- One may examine computational systems and thus indeed learn something about *computation*. That presupposes a criterion that allows us (even if not yet perfectly) to identify *computers* independently of such an examination.
- Adopt the "computational stance" toward a system, interpreting its states and transitions as data structures and algorithmic steps, respectively. Whether this approach pays explanatory dividends, whether it "reveals something important and perhaps unexpected about the real nature of the device", cannot be decided *a priori*, but is part of messy retail business that is empirical science. At any rate, interpreting the device *as* computational presumes an explicit account of the term "computational."

These explanatory schemes are mutually exclusive, but Churchland and Sejnowski's remarks suggest they seek to combine them in an attempt to get around the conceptual requirements as part of both strategies. To see what is going on here, we must get an idea of how the definition of *computation* fits within the larger project of conceiving of cognition in computational terms.

2.4 Reverse Engineering

The computational story about cognition can be exploited in two different ways, roughly corresponding to the classical, symbolic approach on the one hand, and connectionism on the other.

The first strategy postulates semantically *transparent* (Clark) internal symbols more or less directly corresponding to linguistic expressions. An algorithm working

[4] It is hard to see how the complexity of the device's input-output behavior should warrant ascription of the term "computer." The multiplier that appeared in Cummins's and Schwarz's text, for instance, evidently does not implement a "complex and inobvious" function.

on this inner syntax then accounts for the systematic transitions between intentional states. On this view, the brain is just nature's intricate way of implementing the computational architecture underlying thought. In terms of the tripartite definition of "computation" elaborated in the preceding section, the classical paradigm entails comparatively simple algorithms (or at least, that was the conjecture), but highly abstract physical states.

Connectionism, on the opposite, organizes large numbers of fairly simple neurons into networks capable of dealing with elaborate tasks, although in a semantically opaque manner (arguably – at a minimum, the relation to linguistic representations is rather complicated when compared to the direct correspondence suggested by classical theories). Offering a more compelling story from a physiological point of view, the theory falls short of providing an elegant account of linguistic competence and (thus) "high-level" cognition. To put it more carefully: what one will find in a connectionist model are neurons and connections, not sentences and reasons.

Accordingly, "computational" interpretations of connectionism focus on the physiological structure of the brain, often operating with a blunt isomorphism of neural and algorithmic "states." This way of setting up the correlation gives rise to intricate computational procedures which are unlikely to yield to the kind of perspicacious cognitive interpretation found in classical models.

Following to this coarse characterization, the "computational" narrative can be developed into theories that focus on an abstract, phenomenal account of cognition (in case of the symbolic paradigm) on the one hand, and "messy" physiological detail (connectionism and more recent work on so called "dynamic cognitive science") on the other. It has been argued that the latter strictly speaking do not yield to computational interpretations anymore (most notably in what is known as the "Systematicity Debate"), but, as we have seen, the definitions are loose enough as to render both interpretations viable.

The way *computation* is invoked with connectionism is paradigmatic of the kind of *reverse engineering* that pervades cognitive science. Churchland and Sejnowski argued that, in order to license attribution of computational characteristics to a system, the function it *satisfies* (in Schwarz's and Cummins's terminology) would have to be "complex and inobvious." But in some sense, the idea of "discovering" the input-output function associated with a certain device is inconsistent with their definition of "computation", since knowledge of this function figures as a prerequisite in adopting the computational stance. To make this more concise, let me rehearse their definition:

> We count something as a computer because, and only when, its inputs and outputs can usefully and systematically be interpreted as representing the ordered pairs of some function that interests us. Thus there are two components to this criterion: (1) the objective matter of what function(s) describe the behaviour of the system, and (2) the subjective and practical matter of whether we care what the function is.
> **(Churchland and Sejnowski, 1992, p.65)**

The problem, then, blatantly, is this: if we do not even know what function the system is supposed to compute, how can we decide **a)** whether its intermediary states execute an appropriate algorithm, and **b)** tell whether we are even interested in the function we eventually happen to discover? I assume most cognitive scientists would respond by pointing out the *hypothetical* character of computationalist theory – whether the mind/brain can best be explained in these terms is an open problem. Turning the above criterion into a *hypothesis* regarding the computational nature of the brain, it would have to contain – *mutatis mutandis* – one of the following statements:

(a) *we assume that the brain's input-output behavior can be interpreted in terms of a function* ***f***

(b) *we assume that this function interests us*

Quite obviously, the second presumption, regarding the "subjective" part, is nonsensical. The interest we take in the putative function satisfied by the brain can not serve as the content of the hypothesis proper. It is rather the other way around: the concern for whatever function this turns out to be makes one adopt the computational stance. What remains to be seen, then, is whether the first assumption is legitimate – that is, whether it is possible to identify *intermediate states* in relevant parts of the brain that execute an algorithm for this function.

In the kind of research envisioned (and carried out) by Churchland and Sejnowski, computational interpretations are used as a means to reveal the functions performed by certain sections of the brain. For instance:

> Consider, for example, the neurons in parietal cortex whose behaviour can be explained as computing head-centered coordinates, taking positions of the stimulus on the retina and position of the eyeball in the head as input. ... Knowing that some neurons have a response profile that causes other neurons to respond in a certain way may be useful, especially in testing the computational hypothesis, but on its own it does not tell us anything much about the role of those neurons in the animal's visual capacity. We need additionally to know what the various states of neurons represent, and how such representations can be transformed by neural interactions into other representations.
> **(Churchland and Sejnowski, 1992, p.68)**

In order to make sense of it, one must turn the criterion upside down: the computational interpretation is not validated *a priori*, but gains momentum from its utility in explaining the mechanisms accounting for the systems overall behavior. From this perspective, computationalism assumes a more heuristic status, rather than entailing any substantial theoretical commitments. But of course, computationalism is usually regarded to involve precisely the latter.

If computationalism serves as a hypothesis, we must ask ourselves: what would render it false? Along the lines of the first statement I offered above, what would

amount to a refutation is to show that the brain can not be interpreted as passing through physical states corresponding to states in an algorithm. But absent any specific algorithm or function, what does that mean?

As part of the criticism regarding the notion of *computation* put forward above, I argued that what follows from the standard definitions is a kind of *pancomputationalism* – that is, everything has the property of being a "computer" (and that was the case even for more rigid definitions which sought to preserve *objective* qualities). I believe we are now at a point where we can see that this is not merely an unfortunate collateral effect stemming from the tangled conceptual foundations of an otherwise well-defined research program. If I am right, it is unlikely that one could present a definition of *computation* which resolves this issue and leave the rest of the philosophical edifice of computationalism untouched, simply because the fuzziness of the notion plays an important methodological role. Churchland and Sejnowski are committed to a kind of *reverse engineering*, assigning computational models to neural structures which plainly reflect their causal organization. Due to the protean qualities of the concepts involved, in principle it is always possible to cover the bare physical system with this computational layer.

This is in line with Cummins's and Schwarz's characterization of algorithms as "abstract causal processes" (Cummins and Schwarz, 1991, p.63) – the physical system is not analyzed with respect to constraints deriving from an algorithmic model, but instead serves as the blueprint for a crude sort of algorithm. To speak of a computational *hypothesis* in this context is utterly misleading, in that its truth conditions hardly put any restrictions on the devices to which it is applied. As long as the causal interactions governing the system's behavior can be effectively simulated, computationalism (in this sense) is vindicated.

That does not mean that explanations involving representations and computational processes never pay additional dividends when compared to purely causal descriptions. What matters is that for computationalism to be true, according to the texts I have discussed, *they do not have to*. As a consequence, taking the computational stance is completely trivial and cannot "reveal something important and perhaps unexpected about the real nature of the device and how it works." The vacuity of this paradigm is illustrated by Cummin's and Schwarz's commentary on "non-computational" connectionism – the only serious alternative to computationalism they can conceive of is given by the following situation:

> The representational states, while causally significant, are states in a dynamic system whose characteristic function–the function defined by it dynamical equations–is not itself computable.
> **(Cummins and Schwarz, 1991, p.69)**

That is, the computational interpretation does not apply only in situations where the internal causal structure can not be computationally simulated – where it gives rise to an overall behavior only describable by functions that are not themselves computable.

2.5 Cognitive Functions

What if it turns out that our brain is indeed a "computer" in this (rather weak) sense envisaged by advocates of computationalism? What properties might the functions thereby discovered exhibit? Cummins and Schwarz contend that it is in virtue of satisfying these "cognitive functions" that systems can be called "cognitive" at all:

> Systems are cognitive in virtue of satisfying cognitive functions. We think of a system as cognizing a domain rather than merely responding to an environment when that behaviour essentially involves respecting epistemological constraints appropriate to some specified domain, the domain it is said to cognize. We can conceive of these epistemological constraints as determining a cognitive function, and this, in turn, allows us to think of cognitive systems as systems that satisfy cognitive functions. Computationalism is the hypothesis that systems are cognitive in virtue of computing cognitive functions.
> **(Cummins and Schwarz, 1991, p.63)**

These short remarks obviously do not shed much light on how we should conceive of these functions. The notion of an "epistemological constraint" is suited to further obfuscate the issue, rather than clarify it. Given the status assigned to "cognitive functions" by the authors, this lack of information is startling. Instead of engaging in wild speculation as to what they could have in mind, I simply want to point out that, first, the concept is used in a *philosophical* context. Cognitive functions figure in a description of *what* it is to be a cognitive system, while computationalism in turn offers an account of *how* the relevant properties are attained. Accordingly, there could be various ways for systems to satisfy these functions, but their status as distinctly *cognitive* would be independent from this kind of detail.

It thus seems reasonable to suppose that it must be possible to present an analysis of cognitive functions prior to empirical research. However, the project characterized by Cummins and Schwarz appears to be inspired by what I have referred to as *reverse engineering*. Cognitive functions are to be discovered in the course of describing systems in computational terms. The authors attempt to explain this strategy in the following way:

> One does not, for example, set out to build a computational system to play chess with a specification of a chess function (from board positions to moves, say) in hand. On the contrary, the only way we know how to specify such a function is to build a computational system to play chess. When we have built a computational system to play chess, however, we have specified a chess function.
> **(Cummins and Schwarz, 1991, p.71)**

Comparing the task of building a chess computer to that of creating cognitive systems is evidently flawed. Chess represents a well-defined problem domain. Although one may not set out to create a machine capable of playing chess with an

explicit "chess-function" in mind, there exists a crisp understanding of what moves are legitimate, what strategies are desirable, and so forth. There is, so to say, a *theory* of chess. We certainly do not build (or dismantle, for that matter) chess computers to discover the rules of the game.

But as far as the mind/brain is concerned, that seems to be just what Cummins and Schwarz are after. According to their approach, systems are cognitive in virtue of satisfying cognitive functions. Hence, the mind/brain must satisfy such a function. And if computationalism is correct, then we can find out what this function looks like by examining the brains putative computational structure. In the end, or so the story goes, one could thereby arrive at a theory of cognition.

2.6 Summary

The standard definition of *computer* in cognitive science revolves around mathematical isomorphisms coupling physical states with (representational) states in an algorithm. The problem common to different variants of this explanatory strategy is that they easily arrive at an understanding of *computer* that is too inclusive. It often entails what is known as *universal realizability*: *every* physical system implements *every* algorithm. By regimenting the notions involved in the definition, it is possible to avoid this extreme consequence. But even so, there remains a surplus of interpretation: physical systems may implement multiple algorithms, and *every* physical system implements *some* algorithm.

One may attempt to bypass this issue by locating the interest-relative aspect within the content of "computation" itself. Following to this train of thought, the fact that *anything* can be interpreted as a computer does not diminish the utility of computational analysis of the brain/mind. It is rather that the protean quality of the concept is invoked in legitimizing this approach. This, I argued, is methodologically inconsistent: an analysis of the mind cannot yield new insights about the nature of computation unless there are independent criteria fit to justify this *modus operandi*. Conversely, if the status of computational interpretations is to be mainly heuristic, the content of "computation" needs to be well-understood in advance to applying it in non-standard cases.

I tried to account for this inconsistency by making explicit its critical role within computationalism: the notion of *computation* needs to be loose enough to dovetail with processes of *reverse-engineering* the mind. The computational mode of analysis is thereby detached both from considerations of explanatory success *and* questions about whether it is a priori suitable for the object of study.

In this way, cognitive scientists are free to reveal the putative computational structure of the mind/brain in an attempt to specify the (mathematical) function(s) that are thought to constitute cognition. This is paradigmatic of cognitive science as a discipline that seeks to explain *what* cognition is by analyzing *how* cognitive systems work.

3 Representation

This section is concerned with the other of the two concepts constitutive of CRUM, viz. with *representation*. I have already characterized (if rather briefly) in the introduction what explanatory role is assigned to this notion in accounting for some fundamental aspects of cognition. To recapitulate: Representations figure in a scientific explanation of what is referred to as *intentionality* – the fact that thinkers have beliefs, feelings, desires etc., with contents that are typically *about* some object (where "object" is to be understood in a rather broad sense). Individuals act in accordance with these intentional states, or draw inferences from them, changing their beliefs in a way that respects reasons. But how, one might be puzzled, do they accomplish this? After all, reasons are not the kind of causes normally found in scientific theories. Nonetheless, interpreting individuals as obeying to norms of rationality is arguably the most potent way to predict their behavior.

The main idea of cognitive science is to make this intelligible in terms of representations resembling data structures in a computer program. One can then think of the content corresponding to an intentional state as encoded in these structures, and explain the causal efficacy of beliefs by positing computational procedures ("implemented" by the mind/brain) whose operations are sensitive to that code.

In what is nowadays known as "classical" cognitive science, the details of this story are spelled out by positing syntactic representations – instantiated in the brain – corresponding to linguistic expressions. Accordingly, descriptions (in ordinary language) of individuals as intentional succeed in reliably predicting behavior because they are mirrored by internal *symbols* that are part of the physical world.

Following this line of thought, the mystery surrounding the causal powers of intentional states vanishes, offering a compelling materialist theory of the mind. Due to the syntactic quality of the representations employed, such views are thought to postulate a language of thought (LOT). More generically, the representations used in classical theories are what Andy Clark calls "semantically transparent" (Clark, 2001) – that is, their models make direct use of vocabulary that characterizes the problem domain.

But the notion of representation is not limited to these kinds of models. As I have already indicated, it is common to almost all paradigms in the field of cognitive science (with the important exception of recent developments in "dynamical" cognitive science). Most notably, on conventional interpretations, connectionism is perfectly representationalist in spirit.

Given the diversity of approaches and the fact that opposing camps often question the legitimacy of the specific kind of representation employed by the others, it is somewhat hard to track down important features shared by all accounts of the

concept, but, I believe, at least the following two are undisputed:

(a) Representations are discerned by their *functional* properties

(b) Representations are open to *semantic interpretation*

For instance, connectionist representations are numerical vectors corresponding to patterns of activation in a neural network (or abstractions from such vectors). Accordingly, on the one hand, these vectors are just mathematical objects determining the response of the network to any given input. Yet at the same time, they are associated with a semantic content, with something they *represent*. That is to say, they are of two worlds: one purely consisting of functional relations (realized in physical systems), the other rich with such things as meanings and denotations.

The subsequent discussion pertains to the relation of these worlds as construed by various theories in cognitive science. More specifically, I intend to examine each model vis-à-vis the following questions:

(1) Is there a correspondence between the functional and semantic aspects attributed to representations? In other words: do they line up?

(2) What significance is assigned to such a relation? What are thought to be the ramifications of a possible divergence?

As we shall see, one largely finds agreement here across different paradigms in cognitive science. Given the diversity of the field, this is somewhat of a surprise. There appears to be a general consensus regarding the theoretical import of making the causal-functional and semantic stories match. The rationale is that unless one succeeds at properly relating these descriptions, it is impossible to explain the causal efficacy of intentional states. I will argue that this view is flawed, and that there need not be any discernible correlation between causal-functional states and intentional states of an individual for the latter to figure in explanations of its performances.

3.1 The Classical Case

As I have already indicated, the relation between causal-functional and intentional states (and their contents) envisaged by the classical paradigm is a rather straightforward one. With algorithmic procedures operating on internal sentences, their contents, along with their syntactic structure, determine the result of the processing and thus the cognitive system's behavior. This identity of semantic and causal properties is paramount to a project that seeks to vindicate *folk psychology* – our mutual interpretation as having beliefs and acting according to them – by reconstructing it in scientific terms. In their seminal paper on connectionism (Fodor and Pylyshyn, 1988), Fodor and Pylyshyn characterize this aspect of the classical paradigm in the following manner:

> This bears emphasis because the Classical theory is committed not only to there being a system of physically instantiated symbols, but also to the claim that the physical properties onto which the structure of the symbols is mapped are the very properties that cause the system to behave as it does. In other words the physical counterparts of the symbols, and their structural properties, cause the system's behavior. A system which has symbolic expressions, but whose operation does not depend upon the structure of these expressions, does not qualify as a Classical machine ... In this respect, a Classical model is very different from one in which behavior is caused by mechanisms, such as energy minimization, that are not responsive to the physical encoding of the structure of representations.
> **(Fodor and Pylyshyn, 1988, p.9)**

We are already in a position to give an answer to **(1)** for the classical theory: there is a one-to-one correspondence between the causal-functional and semantic properties of representations. And as far as **(2)** is concerned, this connection is blatantly required if one intends to explain the potency of intentional descriptions in terms of their identity to physical symbol structures.

3.2 Connectionism

Commitment to this identity, however, is not just a feature of the classical approach. As already mentioned, most of the work in connectionism is representationalist in spirit, and the idea that representations and the contents associated with them ought to be causally significant for an intelligent system's behavior is part of this doctrine. As a rough sketch, connectionist representations are given by vectors corresponding to patterns of activation across the units of an entire neural network, in contrast to the symbolic structures encountered within the classical theory. They are commonly referred to as *sub-symbolic* and *distributed* as to highlight these distinguishing properties.

What makes a specific pattern a representation of x is that input (appropriately encoded) of x systematically causes the network to exhibit a characteristic pattern of activation. One way of interpreting their *sub-symbolic* quality is to see that their contents often do not line up with the concepts available in ordinary language. That is, although we are able to identify a specific pattern of activation associated with the feature x, x itself might not correspond to any simple concept we are familiar with (as is the case with so-called *microfeatures*). On a more straightforward view, *sub-symbolic* simply means that the models do not directly operate on syntactic structures.

Moreover, contents are encoded by *distributed* patterns that – in general – range across all units of a network, rather than separately identifying single neurons with

a specific meaning.[1]

In some sense, this gives rise only to a negative characterization of the paradigm, in that it focuses on where connectionist representations *differ* from their symbolic counterparts, rather than spelling out what they are independently of such considerations. Quite generally, connectionists are still struggling to articulate their positions in a way that gives rise to a novel paradigm of cognition.

> Nevertheless, nothing has emerged as what could be called *a*, let alone *the*, connectionist conception of cognition. Classical cognitive science says that cognition is rule governed symbol manipulation. Connectionism has nothing as yet to offer in place of this slogan. Connectionism says that thinking is activity in a neural network–taking the human brain to be such a network. But surely not all activity in any neural network would count as mental in any sense.
> **(Horgan and Tienson, 1991a, p.2)**

In spite of the desire to deliver a different theory of cognition, it appears that (in particular) with respect to the issues discussed here, connectionism and classical symbolic cognitive science have much in common. The question of how to conceive of the relation between the computational and neural levels of description marks a dividing line within connectionism. *Implementationalism* seeks to directly instantiate notions from symbolic computation in neural network models, while *eliminativism* questions the legitimacy of such high-level accounts about mental processing and argues they should be abandoned in favor of a precise rendering of the underlying neural interactions.

3.2.1 Implementationalism

In (Fodor and Pylyshyn, 1988), Fodor and Pylyshyn, alongside their thorough criticism of connectionism as a theory of cognition, endorse its utility as a possible implementation theory. On this view, neural networks are simply nature's way of implementing the species of symbolic computation that gives rise to cognition. This position, which seeks to reconcile both theories, has many advocates within the connectionist community. Clearly, by sharing the fundamental tenets of symbolic cognitive science, this paradigm inherits the conception of an isomorphism

[1] On the other hand, if connectionist representations are construed as patterns across microfeatures, single neurons are normally assigned a specific feature. That is, although, for instance, the concept *cup* may correspond to activation across the entire network, there are individual units that can be interpreted as representing *handle*, *round*, etc. whose activation jointly responds to the presence of *cups*. But in principle, isn't it possible to reduce distributed representations to the activation of single neurons? For any given network, we could take the activation of all units as the input for another network which is trained to detect the pattern distinct of *cup*. By adding a single output neuron and merging the two networks, we could then identify this neuron as a representation of *cup*. I believe the most natural approach is to treat local representation as not necessarily opposed to but simply as a special case of distributed representation.

between functional and semantic properties of representations. That is, inasmuch as implementationalists work towards an instantiation of symbolic computation by connectionist means, connectionist representations are just (possibly intricately) encoded versions of classical symbols. Because, according to the classical picture, semantic contents are conferred on representations by their functional role within the system, these aspects trivially align **(1)**.

This *functionalist* account of cognition, as already mentioned, is thought to answer to salient features of mental activity, such as *systematicity* and *productivity* or the *rational relations* governing transitions between intentional states (see Fodor and Pylyshyn, 1988). Accordingly, **(2)** failure in identifying connectionist representations fit to play the designated functional role is thought to threaten not just classical cognitive science, but with it the very prospect of vindicating so called folk psychology, i.e. our mutual ascription of beliefs, desires, etc.

3.2.2 Eliminativism

A position which is – arguably – more genuinely connectionist denies this very possibility. Eliminativists like P.M. Churchland claim that our inner mental constitution radically differs from the structure of propositional attitudes with which we commonly explain our behavior:

> We ... need an entirely new kinematics and dynamics with which to comprehend human cognitive activity. One drawn, perhaps, from computational neuroscience and connectionist A.I. Folk psychology could then be put aside in favor of this descriptively more accurate and explanatorily more powerful portrayal of the reality within.
> **(Churchland, 1989, p.125)**

He suggests that notions from folk psychology which cannot be reduced to this novel vocabulary need to be abandoned, at least in the context of scientific inquiry. Although Churchland offers reasons for questioning the status of folk psychology that do not turn on the impossibility of such a reduction (such as its putative unreliability), this incongruence assumes a central status.

Let me describe the explanatory resources available to connectionist eliminativism in some detail. Churchland seems to endorse the framework provided by CRUM as a whole, referring to activation vectors as *representations* and their processing as a kind of *computation*. And although these representations cannot exhibit propositional semantics, they are nonetheless assigned semantic interpretants. These *prototypes* (corresponding to regions in the space of activation patterns across the hidden units of a network) are conceived as representing crucial features of prototypical situations in a creature's environment, with each (perceptual) input vector "activating" a specific prototype.

Churchland envisions a unified theory of perception and explanation revolving around the ubiquity of prototypes in cognitive phenomena. (The assimilation of perception and cognition is a salient feature of connectionist paradigms in general,

as we shall see in the context with Smolensky's story about the connectionist representation of *coffee*.) Without going into too much detail: depending on what prototype is activated, the creature (system?) will behave differently – that is, they correspond to *functional roles*.

Inasmuch as prototypes constitute a connectionist semantics (and that appears to be what Churchland has in mind), and prototypes coincide with functional roles of representations, **(1)** there is – again – a close relation between these two aspects of representations.

The radical conclusion drawn by eliminativists from the putative incompatibility of explanations in terms of propositional attitudes and precise descriptions of internal mental structures implies their answer to **(2)**: the semantic categories employed within a theory of cognition must be reducible to the (causal/functional) vocabulary of the natural sciences.

3.2.3 The Case for Constituent Structure

In spite of their almost binary opposition, implementationalism and eliminativism do not cover the entire range of connectionist paradigms. There are some who try establish an alternative on middle ground, defending against the latter the viability of computational interpretations without yielding to the former.

In a paper titled "Settling into a new Paradigm" (Horgan and Tienson, 1991b), Horgan and Tienson set out to explore the possibility of an understanding of connectionism that avoids the horns of the dilemma outlined by Fodor and Pylyshyn's influential text (Fodor and Pylyshyn, 1988) – either proving inadequate as a theory of cognition or merely explaining how classical architectures are implemented in neural structures.

> To lay our cards on the table, we believe that an adequate model of cognition requires complex representations with syntactic structure–just as standard cognitive science says–and that cognitive processing is sensitive to that structure. But, we believe, cognitive processes are not driven by or describable by exceptionless rules as required by the standard paradigm.
> **(Horgan and Tienson, 1991b, p.241)**

Ignoring (for the moment) the last sentence, this is an astonishing concession to the classical theory. But where does the "requirement" of representations with *syntactic structure* come from? The authors seek to motivate their idea by means of a somewhat lengthy example involving basketball. A skilled basketball player, they argue, has to be able to take into account an indefinite number of factors which determine the game, such as the positions of teammates or those of players from the opposing team. Faced with complex situations, she has to make decisions within fractions of a second that will affect the outcome of the match. Horgan and Tienson argue that, for one thing, connectionist systems are better equipped to deal with multiple constraints in parallel than classical ones, but more importantly

> the point is that the system that drives decisions on the basketball court is a rich, highly structured representational system, and it is typical in this respect. Any successful development of connectionism, we believe, must incorporate a representational system of this kind. ... At the heart of it is a representation of an evolving scene. This information might be in some sense imagistic. ... On the other hand, some of the information that goes into court decisions is of a sort that would normally be thought of as propositional, that is, as having a propositional, sentence-like structure. ... We need co-reference and co-predication to encode this information. That is, we need repeatable predicates applied to repeatable subjects. We need syntactic structure.
> **(Horgan and Tienson, 1991b, p.244)**

The crucial observation here is that the content assigned to representations is content "that would normally be thought of as propositional." Further, it is the same kind of content that figures in ascriptions of intentional states, in statements of the form "A believes that p" (where p is some such proposition). If the appeal to syntactic structure is to make any sense at all in accounting for this propositional knowledge, the content of an individual (syntactic) representation must correspond to the content expressed by one of those propositions.

I'm aware that Horgan and Tienson, as they say, want to "draw attention to a neglected region in the logical space of views about cognition" (241), rather than present a full-blown theory which occupies this region. Nonetheless, I believe it is legitimate to take a step back and track down some of the difficulties any theory along their lines will face.

For instance, there is no indication of how to make sense of the correspondence between propositional contents and quasi-syntactic representations. How are the respective contents conferred on these representational structures? How do we know that any representation has p as its content and not q? It is obviously not enough for the representation to have the same content as a proposition to be syntactically identical (or similar) to that proposition. As part of their discussion of the classical theory, the authors refer to its rules as "formal", sensitive to the content of representations in virtue of their syntax mirroring semantic relations (248).

In spite of this observation, they do not offer an explanation of how content is thought to be conferred on the representations which figure in their own model. It is my suspicion that connectionist representations are tacitly conceived of as contentful by means of their internal structure or causal history, but this point is as much in need of clarification as it is for the classical theory.

Leaving this issue aside for the moment and taking for granted that they are contentful, what does the relation between their causal-functional and semantic properties look like? As already mentioned, Horgan and Tienson believe that processing must be sensitive to content. This very processing, on the other hand, can be characterized purely in terms of activations and connection weights, in other

words, at a level where semantic content does not enter. The following passage illustrates how the authors intend to deal with this predicament:

> The denial of cognitive rules does not mean cognitive anarchy. ... Processing must be *sensitive* to representational content and structure, a fact that will reveal itself through the existence of many true *ceteris paribus* representation-level generalizations about the cognitive system. Indeed, our belief (strong suspicion at least) is that a fairly extensive range of such soft generalizations must be true of a system in order for attributions of representational content to its inner states to be appropriate at all, and hence in order for theorizing and explanation to be possible at the cognitive level. We also believe that being describable by soft generalizations but not hard rules is characteristic of virtually all of human cognition.
> **(Horgan and Tienson, 1991b, p.249)**

Evidently, for processing to be "sensitive to representational content and structure", representational content must determine – *ceteris paribus* – its result. Because Horgan and Tienson want to explain this at the level of neural structures, that puts constraints on the syntactic representations which instantiate contents. Although, as the authors emphasize, there may generally be infinitely many vectors that correspond to any specific representation-level content, not just any vector will do – the candidates have to undergird representation-level generalizations.

> It is characteristic of connectionist representations that they permit multiple realizations in ways that can affect the outcome of the processing. The RI[2] rules are what we might call "upward obligatory." Given a node level description of the system, the RI rules determine a representation level description. But in general, connectionist RI rules are not "downward obligatory." Given a representation level description, the RI rules place constraints on permissible node level descriptions, but they may be compatible with many, perhaps even infinitely many, node level descriptions.
> **(Horgan and Tienson, 1991b, p.254)**

In answer to the twin questions I have posed at the outset, in the case of Horgan's and Tienson's theory, **(1)** the causal-functional properties of representations determine their semantic content, and conversely, any specific content can be realized only by members of a particular class of activation vectors.

As far as **(2)** is concerned, such a close relation is required to support *ceteris paribus* representation level generalizations, which in turn "must be true of a system in order for attributions of representational content to its inner states to be appropriate at all, and hence in order for theorizing and explanation to be possible at the cognitive level" (249).

[2] *Representation Instantiation*

Horgan and Tienson (251) mention work by Smolensky as an actual attempt of encoding complex (syntactic) structure in connectionist architectures. Employing an arsenal of mathematical tools (in particular, *tensor product representations*), he intends to show how neural networks can be seen as processing input depending on its compositional structure. As was the case with Horgan and Tienson, one of the main motivations behind these efforts is to respond to Fodor and Pylyshyn's critique of connectionism.

In a paper directly addressed to them (Smolensky (1991)), he attempts to derive a notion of compositionality from the encoding of distributed representations in terms of microfeatures. Along these lines, *cup with coffee* might be represented in a connectionist model as a subset of features such as *hot liquid, upright container, glass contacting wood* etc. (291). *Cup without coffee* would give rise to a different pattern, with the features responding to coffee inactive.

Smolensky argues that these representations have compositional structure in the following sense: by taking the set of features corresponding to *cup with coffee*, and removing from it those present in *cup without coffee*, we arrive at a connectionist representation of *coffee*.

> So what does this procedure produce as "the connectionist representation of coffee?" ... we have a burnt odor and hot brown liquid with curved sides and bottom surfaces contacting porcelain. This is indeed a representation of *coffee*, but in a very particular context: the context provided by *cup*.
> **(Smolensky, 1991, p.293)**

As Smolensky concedes, one can speak of *compositional structure* here only in an approximate sense (293). Because these representations (and that, I presume, is supposed to be the case for connectionist representations in general) are situated in a specific context, the result of this operation is differs depending on context.

> The point is that the representation of *coffee* that we get out of the construction starting with *cup with coffee* leads to a different representation of *coffee* than we get out of constructions that have equivalent status a priori. That means if you want to talk about the connectionist representation of *coffee* in this distributed scheme, you have to talk about a *family of distributed activity patterns*. What knits together all these particular representations of *coffee* is nothing other than a *family resemblance*.
> **(Smolensky, 1991, p.294)**

(I cannot help but think of the appeal to *family resemblance* here as a desperate attempt of covering up the obvious shortcomings of this approach. Ironically, starting from what Smolensky calls *a crude, nearly sensory-level representation* (292) of a cup of coffee, one arrives at something that *does not look anything like coffee*. If one literally removed the parts corresponding to the cup from the impression of a

cup of coffee, what remains is indeed "burnt odor and hot brown liquid" etc., but I doubt that qualifies as *coffee*. The idea seems even more absurd in the cases of *man with coffee* and *tree with coffee* (294).)

Smolensky maintains that while it is legitimate to interpret vector representations as exhibiting compositional structure, such decompositions are in general neither unique, nor do they *directly* determine processing:

> The processing of the vector representing *cup with coffee* is determined by the individual numerical activities that make up the vector: it is over these lower-level activities that the processes are defined. Thus the fact that there is considerable arbitrariness in the way the constituents of *cup with coffee* are defined introduces no ambiguities in the way the network processes that representation–the ambiguities exist only for us who analyze the model and try to explain its behavior.
> **(Smolensky, 1991, p.296)**

The upshot of these observations is that notions from symbolic computations are useful in a higher-level analysis of the behavior of neural networks, but they capture the processing only in an approximate way. As far as **(1)** is concerned, semantic interpretations *only approximately* resemble the causal-functional interactions at the level of neurons.

As I have already mentioned, Smolensky's paper is a response to Fodor and Pylyshyn – he even concedes that his entire work on how to represent complex structures is an attempt of dealing with the criticism put forward by these authors (304). In order to decide where his ideas fit in with the current context – that is, what answer to **(2)** they imply – we have to put them into this larger perspective.

In (Fodor and Pylyshyn, 1988), it is argued that *compositionality* is a defining feature of thought, one that must be accounted for by any proper theory of cognition. Fodor and Pylyshyn claim that connectionism is unable to meet that requirement, in contrast to the symbolic paradigm.

Smolensky, while rejecting their argument as a whole, concedes that the demand for compositionality needs to be taken "quite seriously" (Smolensky, 1991, p.288) – it is among the "nonformal cognitive principles" (300) he thinks connectionist models have to instantiate.

According to what Smolensky refers to as *PTC* ("Proper Treatment of Connectionism"), the relation of these computational abstractions to connectionism is one between macro- and microtheory of cognition. It is part of connectionism's agenda to explain how these higher-level phenomena emerge out of basic interactions at the neural level.

The bottom line is that unless one can translate these principles located at the cognitive level into a story rendered in the language of connectionism, the latter falls short of constituting a theory of cognition.

3.3 Summary

At the outset of this section, I proposed two questions to assess the relation between semantic and functional features of representations that figure in cognitive scientific theories. I tried to show that, across competing paradigms, there is consensus regarding both. Commitment to the identification of internal structures that could undergird ordinary semantic categories appeared as the common denominator for all but one of the candidate theories. In the case of connectionism, the underlying worry is that

> if there is nothing in the actual character of states in connectionist systems that undergirds our intentional attributions, then some critics will see connectionist systems as failing to account for a basic characteristic of cognitive states. Searle makes the same objection to traditional computational accounts of mental functions such as those captured in traditional AI programs, but in at least Fodor's (1975) version of the computational theory, the syntactic operations of the program are supposed to mirror the semantics.
> **(Horgan and Tienson, 1991a, pp.33-34)**

Eliminativism, while denying the existence of such mental tokens, affirms the import of making the semantic and causal/functional stories match by drawing radical conclusions from the putative incongruence of internal mental structure and regular propositional attitudes.

There is reasonable evidence that eliminativism is right in rejecting the idea of an isomorphism linking these levels of description. Do we have a choice but to concede that folk psychology should be abandoned? I believe we do. We merely need to take seriously the philosophical lesson that intentional states are the result of linguistic interaction, rooted in an interpretative practice. Once we have correctly identified their status, attempts of locating beliefs "in the head" will strike us as absurd.

4 Cognitive Science and Functionalism

Consider the way in which Turing, in "Computing Machinery and Intelligence" (Turing, 1950), avoids addressing questions about what it means to "think" directly. Instead of facing the challenge to elucidate the content of the concept, in its place he proposes the conversational test procedure we now know as the *Turing Test*. Surely, something feels wrong about this explanatory strategy. For it is not as if he was not interested in what "thinking" means. On the contrary, the point of his detour appears to be precisely to answer this question.

I think his move can serve as an analogy to the explanatory scheme that is pervasive in contemporary cognitive science: the attempt of explicating *what* it is by providing theories of *how* it is done. In terms of the subject matter: to understand what thought really is, we supposedly require insight into the mental processes – processes ultimately to be reduced to activity in the brain – that form its (physiological) basis.

There is a grain of truth in this: accumulating knowledge about the brain and gaining a deeper understanding of physiological substrate that cognitive faculties rely on will alter the way in which we perceive our own intelligence. But theories in cognitive science have a strong reductionist slant – that is to say, they presume that once we have an idea of how the brain works (on various levels of abstraction), that will essentially be all there is to say about thought (at least in scientific contexts), once and for all answering to the question "What is intelligence?".

I believe this view is fundamentally mistaken – it thoroughly mixes up two issues that need to be dealt with separately for cognitive science to emerge as a well-defined research program. In particular, a thorough philosophical explication of the cognitive capacities around which the field revolves is called for. The (broadly empirical) investigation of how individuals think needs to be augmented with a theory that explains *what it means to think*. To put it in a more technical jargon: the *explanandum* needs to be properly analyzed in order to tell whether the *explanans* does what it is supposed to do.

Specifically, what is required is a (different) theory of *concepts*.[1] The observation that often serves as a starting point for cognitive scientific theories, viz. that individuals have beliefs or thoughts (generally: intentional states), act according to them and adjust them in a reason-governed manner can only be made intelligible against the background of concept-use. Cognitive science needs to take seriously

[1] Chapter 5 is devoted to the presentation of such a theory.

the idea that our cognitive apprehension of the world is mediated by concepts.

Currently, this issue, if it does come into view within the discipline at all, is quickly usurped by the computational-representational framework that constitutes its methodology. Thus, the exercise of explicating the nature of intentional states is transposed into the problem of coming up with a working computational model of cognition. The corresponding debates revolve around whether cognition should best be understood by means of symbolic or connectionist paradigms (or yet other approaches). But the *functional* (algorithmic) states that are invoked as explanatory primitives, by whatever particular cognitive architecture they are instantiated, cannot play the role of beliefs or thoughts for principal reasons.

The details of this criticism will be worked out below. That involves tasks roughly corresponding to at least two different levels of analysis:

(a) Providing evidence for the lack of conceptual analysis as far as the field's explanandum is concerned and illustrate the entanglement of empirical and conceptual issues.

(b) Tracing the roots of the problem to the computational-representational approach (CRUM) taken by cognitive science, and demonstrating how it both reflects and supports this philosophical lacuna.

These issues are interrelated and therefore cannot be dealt with in a strictly consecutive order. What I call a *lack of conceptual analysis* is normally not perceived as such within the discipline because the computational paradigm on which it rests renders the distinctions between the conceptual and empirical underpinnings of cognition vacuous. A criticism of the computational approach in turn provides no satisfying answer unless one can bring into view a different conceptual framework. This will be the main subject of the next chapter.

4.1 "Mindfulness"

Standard introductory textbooks on cognitive science, while covering as many paradigms as possible, at the same time often display a remarkable lack of serious conceptual analysis regarding the cognitive capacities these paradigms are to account for.

Paul Thagard, for instance, hardly goes beyond the trivial observation made at the outset of his text that

> every day, people accomplish a wide range of mental tasks: solving problems at their work or school, making decisions about their personal life, explaining the actions of people they know, and acquiring new concepts like *cell phone* and *Internet*.
> **(Thagard, 2005, p.1)**

Thagard claims that cognitive science is committed to developing models for the entire range of mental phenomena (in a piecemeal fashion, I presume).

This ad-hoc approach may in part account for the absence of comprehensive philosophical groundwork – if cognitive capacities are addressed individually, the corresponding conceptual expositions need not extend beyond the scope of any specific problem. However, even for such restricted domains, the conceptual analysis that comes with the detailed presentation of the particular computational model is reduced to a bare minimum. Assumptions about the nature of intelligence largely remain tacit and are only introduced indirectly, in the context of evaluating particular theories. Thagard proposes several criteria of adequacy, as shown in the list below.

(1) Representational power
(2) Computational power
    a) Problem solving
        (i) Planning
        (ii) Decision
        (iii) Explanation
    b) Learning
    c) Language
(3) Psychological plausibility
(4) Neurological plausibility
(5) Practical applicability
    a) Education
    b) Design
    c) Intelligent systems
    d) Mental illness

(Thagard, 2005, p.15)

Against the background of the present work, it should be noted that the arrangement indicates no principal distinction between aspects (1) and (2) (which roughly correspond to the level of *conceptual* analysis) on the one hand, and (3) to (5) (which can be identified with broadly *empirical* issues) on the other. Furthermore, the conceptual relations pertaining between the selected features remain unexplored. Quite generally, Thagard hardly touches on the philosophical problems raised by the computational-representational approach to cognition his book promotes. Reflections on the concept of cognition are simply beyond the scope of his book.

Does this deficit mark a blind spot in cognitive scientific theory, in the sense that it simply has not been noticed? Of course, this answer would to be too simplistic. It is rather that the *computational models themselves* and the formalization they

involve are intended to take the spot of philosophical theorizing about the nature of intentional states.

Another popular introduction to the field has been put forward by Andy Clark in his book titled *Mindware* (Clark, 2001), a text that is concerned specifically with the *philosophy* of cognitive science. The introductory section illustrates the short route that leads from an informal understanding of intentional states to computational paradigms. Starting from a vague and tentative rendering of what he refers to as "mindfulness," a rather brief story about the explananda of cognitive science (Clark, 2001, pp.1-5) is offered, but only to arrive at what he perceives as the fundamental difficulty the discipline faces. Acknowledging how reasons mediate the transition from one belief to another, or from believing to acting, the dilemma he points out is this: *how can reasons be causes*? On the view of cognitive science he characterizes, the discipline is committed to a fully materialist account of mental phenomena, and that apparently entails translating (or rather: *reducing*) reasons and beliefs to brain-states, whose causal powers can be explained by broadly physical theories.

I have already outlined the classical way out of this dilemma, and cited Jerry Fodor as its most ardent proponent. Fodor maintains that the nomological (and rational) relations obtaining between intentional states should be explained by appealing to a computer model – on this view, beliefs derive their causal powers of systematically bringing about other beliefs and eliciting behavior from their *physical* realization in the flesh-computer that is the brain.

In spite of Clark's explicitly philosophical agenda, the principal problem concerning the missing distinction of explanans and explanandum persists. Questions regarding the nature of intentional states are discussed exclusively in connection with whether they can be made intelligible computationally. Debates revolve around broadly computational models of cognition and the philosophical problems they imply, rather than an independent account of cognition that precedes these models. Instead of taking the indisputable phenomenal reality of intentional states as a benchmark for theories in cognitive science, it is called into question on the basis of the latter's failure to provide an adequate understanding of mental states.

I now want to turn to the discussion of a text that, in contrast to the aforementioned books, puts forward an attempt at elucidating the explanandum of cognitive science. An analysis of its limitations will bring out into the open some of the tacit principles guiding cognitive science.

4.2 Inference and Intelligence

Robert Cummins's account of cognitive capacities in *The Nature of Psychological Explanation* (Cummins, 1983) bears witness to the distortion of empirical and conceptual questions that is pervasive of cognitive science. In an attempt to explicate what is distinctive of cognition, he oscillates between both aspects, applying, as it were, normative standards of rationality to the internal causal structure of cognitive systems which is ultimately bound to fail this test upon closer examination. The

details of this predicament illustrate the ramifications of not separating both types of questions, which is why I will take a closer look at Cummins's text below.

Early on, he boldly claims that "until quite recently, no one had the slightest idea what it would be like to scientifically explain a cognitive capacity" (52). Leaving aside the matter of whether his rejection of "philosophical psychology" is justified, let us focus on the broadly computationalist conception he proposes in its place. In a nutshell, he maintains that cognition should be modeled on "epistemologically constrained sentence transformation functions," or, quite simply: inference. According to this story, cognitive capacities are just *inferentially characterizable capacities* (*ICC's*) (53). However, as Cummins hastens to add, not all ICC's are cognitive capacities, since "cognition is a propositional attitude, and certainly the exercise of an ICC needn't be a propositional attitude" (53-54). In consequence, he resorts to using the term "cognition*" when referring to the outputs of ICC's, postponing the analysis of its relation to *cognition* proper in favor of an investigation into the properties of ICC's. Interpreting performances as the results of exercising an ICC involves two criteria, Cummins claims:

(a) "First, we must understand what make the output right or cogent relative to the input (i.e., *we* must be able to infer outputs from inputs)" (54).

(b) "Second, we must see that the system is so structured as to (characteristically) exploit in producing an output whatever it is that makes the output cogent relative to current input" (55).

In some sense, (a) certainly seems too strict. In order to (correctly) interpret someone as undertaking an inference, we need not be inclined to follow the inferential pattern we ascribe as part of that interpretation ourselves. Even in cases where outputs/conclusions are subject to assessments of correctness (relative to inputs/premises) on part of the interpreter, Cummins insists, unless the interpreter is able to reconstruct the complete inferential pattern leading to the output, the interpretant cannot be attributed with an ICC. This blatantly runs counter to the fact that we commonly treat performances as inferences without having an exhaustive understanding of the underlying inferential pattern or reason to believe it is correct.

(b) ultimately boils down to identifying appropriate "physical transitions" as what corresponds to "drawing the right conclusion" (55). Cummins struggles to give a clear sense to this second requirement, and it is certainly most controversial. The question that immediately comes to one's mind is: what if *we* (our brains) do not satisfy this criterion? That is, if, at a physiological level, no transitions can be identified that "correspond" to *rational* relations, should we stop thinking of ourselves as *inferring*?

Putting this issue aside for the moment and continuing with the discussion of Cummins's twin criteria for ICC's, we are immediately confronted with their inadequacy. (a) and (b), jointly licensing *inferential analysis*, were intended to (ultimately) give rise to standards of intelligence. But (b), in particular, falls short of

doing this job. (b) was meant to explain how the system "gets it right," in contrast to mere "instantiation." But faced with a device that merely combines elementary logical gates (56), Cummins admits that it cannot be characterized as intelligent (or as "figuring out" the correct results), at the same time conceding that he doesn't "see how to argue the point." In an attempt – nonetheless – to provide necessary conditions as to tighten his grip on the concept of *intelligence*, he introduces the notion of "informed choice" (57). Even though he "can't quite say what informed choice is" (57), he illustrates his idea as follows: "If the process leading to output can be adequately flow-charted without branches, or if the branching is totally insensitive ... to factors bearing on the correctness of the output, then the output is not intelligently produced" (57).

Accordingly, Cummins intends to focus on the class of sophisticated ICC's involving "interlocking informed choices" (58) corresponding to a *rationale*. He claims that "such a capacity is explained only on the hypothesis that the system executes the rationale" (58). But what does "execute" mean here? Let I be a causal mechanism ("symbol cruncher") *instantiating* a capacity C characterizable (only) via a rationale R – then I must be *isomorphic* to R: "I is R in disguise" (59). Let us assume that, indeed, there are criteria allowing us to determine whether a system is isomorphic to a given rationale. What, Cummins wonders, do we make of systems (instantiating I) that exhibit C without being isomorphic to a rationale R we thought was necessary to explain C? These amount to "debunking" (60) discoveries, in the sense that what was envisaged as an intelligent capacity to be accounted for by an isomorphism involving a rationale R turns out to be result of brute "instantiation." If such a characterization is possible, Cummins concludes, rather than explaining intelligence, it merely "explains away the appearance of intelligent cognition" (61).

At this point, I want to bring to an end the detailed discussion of Cummins's text and develop an understanding of how it fits into the bigger picture of the present investigation. In an attempt to characterize distinctively *inferential* capacities (ICC's), he – correctly – identified the dimension of *justification* they necessarily involve. Simply producing a sentence in response to an input sentence is not enough: it must be possible to for the result to be assessed with respect to standards of *correctness*. To use Cummins's phrase: it is a matter of "getting it *right*." Unless one can reasonably ask the question of what *rational* relations link inputs to outputs, the capacity in question cannot be described as inferential. But Cummins immediately distorts this insight into the *normative* fabric of inference by linking it with questions regarding the etiology of the response. He demands that, for a system to exhibit an ICC proper, it must be "so structured as to (characteristically) exploit in producing an output whatever it is that makes the output cogent relative to current input" (55) (requirement (b) discussed above). Determining whether a system can be interpreted as inferring thus turned into a matter of telling whether its causal structure is "isomorphic" to a rationale (the latter representing the normative dimension of inference).

It is this correspondence between the normative and causal dimensions, Cummins

argues, that warrants ascription of genuine inferential capabilities and therefore intelligence. As a consequence, we might discover, upon closer examination, that a device we thought was intelligent (in the sense of instantiating a rationale) realizes a capacity by "dumb" instantiation. Cummins mentions Schwartz's "density computer" as an example: instead of calculating the density of a sphere and comparing it to 1 (executing a rationale), it simply "measures" it by testing whether it sinks if put into a bucket of water (60). The putative upshot is that one can distinguish intelligent systems from nonintelligent ones by their instantiation of rationales.

This demarcation, however, is utterly fragile. After all, what if we simply turn the requirement on its head: if a (presumably inferential) capacity is realized by a system, what keeps us from simply "reading off" a rationale from its causal structure? Doesn't the explanation of how the "density computer" arrives at its result provide the kind of justificatory relations characteristic of a rationale? Even though the density is not *calculated*, the device correctly identifies objects for which it is greater than 1. And this is no coincidence: if the "input" sinks, that provides sufficient reason to make the corresponding inference. In other words: the system "gets it right" by exploiting facts that "make the output cogent relative to current input." One cannot characterize the system as "unintelligent" on the basis that it fails criterion (b), simply because it doesn't. A rationale, an algorithm whose transitions are subject to assessments of correctness, cannot literally be part of a physical system. Just as the definition of *computation* centered around isomorphisms with an algorithm proved too weak to pick out the class of computational systems, the current one cannot generate the desired dividing line.

Let me now pick up the question previously left unanswered: are "debunking" discoveries possible even in the context of human cognition? Following Cummins's account, we would have to give a positive answer. If it turns out that the workings of the human brain can be accounted for without recourse to rationales, that would amount to "explaining away the appearance of intelligent cognition" (61). But that is to explain away the explanandum itself!

I believe the only reasonable conclusion is to drop criterion (b) and insist on the conceptual gap that separates normative and descriptive discourse (see Shanker, 1998, Ch.1, §6). Although Cummins is on the right track in pointing out the essential normativity of inference, he is wrong in projecting processes of justification onto the causal structure of cognitive systems themselves. Indeed, why should the physiological processes in someone's brain that can be identified as the cause of a specific response figure in assessments of whether it was *inferred* from a set of premises? What is required is not that the response is the result of mechanisms that can be seen as the instantiation of strategies of justification, but that it *can be justified* (by presenting suitable premises, for instance) under certain conditions. Cummins's treatment of the inherently normative dimension of inference (and cognition) is paradigmatic of cognitive science by blurring the line separating it from activities in the brain that belong to realm of the natural sciences (possibly in an attempt to reduce the former to the latter).

But looking for rational relations as if they were literally part of the cognitive

systems under scrutiny is entirely futile. At the level of physical descriptions, one can only describe causal interactions – questions of whether individual activities are *correct* simply do not arise, even if one can identify the process as the implementation of a rationale (an algorithm) that *is* subject to normative assessments of this kind. On such a description, intelligent systems are always "dumb." An analysis of their causal makeup only "offers exculpations where we wanted justifications," as John McDowell (McDowell, 1994, p.8) puts it (albeit in a slightly different context).

Merely pointing out the distinction between the normative and causal is unsatisfying, however. Even though it is certainly impossible to reduce one to the other, we must not conclude that they are completely unrelated. Questions, for example, pertaining to the structure of our brain, do, of course, have great significance in the endeavor of understanding our cognitive capacities. But we must assign these issues their proper place within an understanding of the mind by insisting on the primacy of the normative in specifying the explanandum. We must first grasp – in normative terms – what it *means* to think before we can begin to ask ourselves *how* it is done. The relation between normative and causal descriptions accordingly has do be redefined: *How do systems achieve the capacities required for them to be (correctly) interpreted as engaged in the normative practices that are distinctive of cognition?*

4.3 Functionalism

As I have mentioned above, computational theories offer an understanding, for cognitive scientists, of how intentional states can be physically "real" in a way that explains their causal efficacy. We have the mental states we do because they correspond to algorithmic states implemented by our brains, and we reason the way we do because rational transitions mirror an internal program – or so the story goes. But this story can also be read backwards: not just as an account of how beliefs are effective in human cognition, but as a theory about the nature of beliefs.

This philosophical stance is known as *functionalism*, and I will address it now in more detail. In a nutshell, it proposes the identification of mental (intentional) states and states in a computational system. Following this line of thought, the belief that *there is large number of dogs in the neighborhood*, instead of being discerned by intrinsic properties, derives its specific relevance (and content) from the functional role it plays within a larger system of beliefs, that is, by giving rise to or deriving from certain other beliefs in various ways. Originally, functionalism was motivated by the insufficiency of a straightforward identification of mental states with the chemical/physical (brain-) states out of which they emerge. The problem is solved by moving to a different level of abstraction: many physically different systems can be instances of the same functional system, just as hammers can be made out of different materials while sharing a function (driving in nails). This distinction between the functional (computational) system and the physical system which instantiates it gives rise to a conception of mental states that is attractive to

cognitive scientists because it seemingly avoids dualism without regressing toward crude materialism.

Functionalism provided the philosophical foundations of the classical symbolic paradigm, and it appears to form the (tacit) underpinnings of the computational-representational understanding of the mind contemporary cognitive science promotes. If intentional states are construed as states in a computer system (algorithm), the task of the empirical subdisciplines boils down to finding the specific algorithm that *is* cognition. I believe it is possible, from the vantage point of functionalism, to account for the methodological idiosyncrasies I have tried to identify.

First and foremost, the distinction between conceptual analysis of thought and computational modeling I have been advocating collapses or becomes irrelevant. If intentional states *are* computational states, any explication of the conditions under which a system can be perceived as deploying cognitive capacities is *essentially* computational. In other words, the computational framework is not imposed onto cognition from outside, but is thought to reveal its very core.

Functionalism also legitimizes the otherwise methodologically questionable identification of computational models with psychological theories. If a computational model reproduces the behavior observed in psychological experiments, chances are it is isomorphic to the functional system underlying the mental faculties under examination.

4.3.1 Computation and Representation

Some of the issues discussed in context with the key concepts of *representation* and *computation* appear more clearly against the background of functionalism: I argued that the semantic dimension of *representations* is typically cashed out in terms of their functional role. More precisely, the causal efficacy of semantic content is explained by invoking the functional properties of the representations with which that content is associated. But the relation is perfectly reciprocal: semantic content is conferred on representations by their functional features. In this sense, different paradigms exhibit an implicitly functionalist theory of semantic content. But the contents associated with representations is *conceptual* content, which is ultimately to account for or warrant the ascription of intentional states to an individual. Hence, objections raised against functionalism clearly bear on these theories.

Further, I indicated how theorists struggle with determining criteria that decide whether a system should be regarded as *computational*, as implementing an algorithm. I stressed the fact that the typical definitions fall short of establishing sufficient (or necessary) conditions. While this is certainly not a knock-down argument against computational interpretations of the brain/mind in general, it discredits attempts of establishing a level of computational description prior to closer empirical examination. These definitions of *computation* (implementation) play a crucial part in arguing that, *in principle*, the mind is a computer. Yet, they could be incorporated into arguments proving the computational nature of virtually anything.

There several ways to read this strong commitment to computationalist renderings of cognition. On the one hand, it is hard to deny the abstract elegance of combining an interpretation of thought as symbol manipulation and the Church-Turing thesis in a vindication of computationalism. Furthermore, the prospects of AI increase with the cogency of computational theories: although, at this time, it may not in practice be feasible to replicate the – potentially complex – computational processes occurring in the human mind on ordinary computers, it must certainly be possible. This entanglement of cognitive science and AI is well illustrated in Stuart Shankers characterization of Turing's philosophical project:

> With all of the rhetoric about the computational possibilities being opened up ... it was easy to overlook the fact that, in order to defend his philosophical thesis – viz., his proof that, if not quite yet, at some point in the future, machines will indeed be capable of thought – Turing was led deeper and deeper into the development of an appropriate psychological theory: viz., that *thinkers compute*. By the time he came to write 'Computing Machinery and Intelligence' he was explaining how his real goal was that of 'trying to imitate an adult human mind'
> **(Shanker, 1998, p.37)**

I want to focus on a different aspect, however. The computational paradigm of cognition is inclined to move from an account of ordinary intentional states to a more fine-grained description of subconscious processes. One of the reasons behind this strategy is that the behavior of individuals is "underdetermined" by ascriptions of such standard psychological states – not even an exhaustive characterization of someone's state of mind suffices for precisely predicting what someone *will* do,[2] let alone provide the resources for devising an algorithm that *explains* why certain behavioral patterns were elicited.

One way of closing these gaps is to fill in hypothetical steps performed on a subconscious level. This idea more or less represents a top-down explanatory strategy, but it can be exploited in the opposite direction as well: rather than positing mental processes on the basis of what is missing in conscious reasoning and assuming that they are implemented by the brain, one can start by directly interpreting brain processes as computing functions in an attempt to account for "higher-level" cognitive activity in terms of such basic operations.

Either way, these theories are conceived as closer to the reality of the mind than ordinary psychological descriptions, and if there is anything to the latter, they need to be reconstructed out of the algorithmic building blocks provided by the former.

[2]The fallibility of predictions based on the attribution of mental states drawn from the arsenal of ordinary beliefs, desires etc. was cited in the previous chapter as one of the reasons for Churchland to question the legitimacy of *folk psychology*. However, as we shall see in the next chapter, the primary role of intentional states is not their figuring in accounts of an individual's behavior. Instead, they should be conceived as *normative* states – pertaining to what somebody *ought* to do – that are instituted on individuals by social practices independently of their use in *causal* explanations.

The functionalist premise that psychological states must be understood as computational states is retained even when the conceptual resources render translations between ordinary intentional states and these algorithmic states unlikely. That is, rather than letting go of the idea that intentional states can be modeled in terms of functional roles, the very conception of intentional states (and cognition quite generally) is accommodated as to fit the functionalist scheme. It is in this context that Clark refers to connectionism as a kind of "microfunctionalism":

> The functionalist, you will recall ... , identifies being in a mental state with being in an abstract functional state, where a functional state is just some pattern of inputs, outputs, and internal state transitions taken to be characteristic of being in the mental state in question. ... Now imagine instead a much finer grained formal description, a kind of "microfunctionalism" that fixes the fine detail of the internal state-transitions as, for a example, a web of complex mathematical relations between simple processing units. Once we imagine such a finer grained formal specification, intuitions begin to shift. Perhaps once these microformal properties are in place, qualitative mental states will always emerge just as they do in real brains? ... it does not strike me as crazy to suppose that real mental events might ensue. Or rather, it seems no *more* unlikely than the fact that they also ensue in a well-organized mush of tissue and synapses!
> **(Clark, 2001, p.36)**

Hence, the functionalist foundations of cognitive science offer an explanation for the shift towards "microtheories" of cognition, with an increased focus on disciplines such as neuroscience. Even recent developments united under the banner of so called "dynamic" cognitive science, whose theoretical agenda can be characterized as explicitly anti-representationalist and anti-computationalist, can be interpreted as following this route by increasing the level of detail to a point where the original framework becomes obsolete.

The bottom line is that various paradigms assembled under the banner of the computational-representational understanding of mind can be interpreted as *variations of the functionalist theme*. Now, I do not want to claim that this amounts to a deep or revolutionary insight, given that the terms *computationalism* and *functionalism* are sometimes used interchangeably, and cognitive scientific theories are overtly construed as functionalist theories.

However, I hope this observation can shed light on the confusion of, as I put it at the outset, *explanandum* and *explanans* that is prevalent in cognitive science (in case of the last paragraph: mental states and their physiological causes). Against the background of functionalism, as I have argued, this distinction is blurred – the idiosyncrasies I have presented and functionalism are inextricably intertwined, and one must first cut this Gordian knot before a new conception of cognitive science can emerge.

4.4 What Functionalism can't do

Fortunately, thoroughgoing criticism of functionalism has been put forward by a number of distinguished philosophers. Here, I want to focus on the lucid rejection of functionalism advanced by one of its former proponents: *Representation and Reality* by Hilary Putnam (Putnam, 1988). The subsequent paragraphs present some of the central objections worked out in this book.

4.4.1 Semantic Externalism

The first strand of thought which leads to the dismissal of functionalist theories starts with what is known as *semantic externalism*. Putnam illustrates this phenomenon by appealing to the following ("Twin Earth") scenario (Putnam, 1988, pp.30-33): we are invited to imagine a planet quite like ours, inhabited by creatures that are identical to us in almost every respect except for one detail – their physiology is sustained by a liquid that, even though they refer to it as "water" and it is indiscernible from H_2O, is really some other substance, XYZ. Indiscernible, that is, by means available to us or the people of Twin Earth at the time of 1750, prior the development of modern chemistry on both planets.

Now, even though these individuals are unable to distinguish H_2O and XYZ, the term "water" has a different referent on Earth and Twin Earth, respectively. And not just from the perspective of individuals who *are*, in fact, in a position to tell apart these substances. Putnam argues that an inhabitant of Earth (who, by cosmic coincidence, ended up on Twin Earth), had she referred to XYZ as "water," would have made a *mistake*, not because her *conception* of "water" was different (it wasn't), but because the *stuff* is different.

Intuitively, most of us would probably be inclined to perceive this difference in *reference* also as a difference in *meaning*. As a consequence, we can imagine two specimens from the above planets whose "mental representations" are completely identical, who nonetheless are in different intentional states because their thoughts are *about* different objects.

The upshot of this entire scenario is that a brain's functional organization cannot in itself determine the content of beliefs attributed to its owner, because the environment makes an irreducible contribution to that very content. One simply cannot individuate meanings "in the head," in isolation from the social and nonsocial environment.

4.4.2 "Narrow" Content and "Broad" Content

In an attempt to lessen the sting of this argument, Putnam suggests, one may propose to distinguish two kinds of semantic content, "narrow" and "broad." Working with these notions, one can then claim that while narrow content does not determine broad content (including reference), narrow contents play an important part in theories of *meaning*. Putnam considers specific conceptions of narrow content

(Putnam, 1988, pp.43-56) according to Fodor (perceptual prototypes) and Block (conceptual role). For both of these theories, he convincingly contends that questions regarding sameness of narrow content are completely irrelevant to questions about sameness of meaning:

- Perceptual prototypes (conceived in connectionist terms) need not even exist for the majority of concepts, and may vary considerably from individual to individual. In particular, they are unlikely to be preserved by translation, in contrast to meanings.

- Conceptual roles, thought of as the role a concept plays in inferences, can similarly undergo radical transformations without changes in the associated meanings. For example, our contemporary beliefs about water are substantially different from what the ancient Greeks knew about water, and therefore the inferences involving the word "water" endorsed by them tremendously diverge from ours. Nonetheless, it would strike most of us as unnatural to interpret this as variations of *meaning*.

4.4.3 A Theory of Reference

There appears to be a straightforward remedy to these issues: instead of confining the scope of functional descriptions to the organization of an individual's brain, one could simply include the environment in definitions of mental states. In order to do that, purely functional vocabulary might not be sufficient. But a combination of functional and physical descriptions that pick out specific situations where an individual is in an intentional state, including possible referents, could do the job, Putnam suggests. Thus, one may attempt to meet the objection that meaning depends on reference, which is partly determined by things "outside the head," without having to give up functionalism.

However, even if we are ready to concede that possibility *in principle*, it is crucial to be aware of the complexity such a theory of reference involves *in practice*. Putnam points out that ascriptions of meanings to tokens proceeds simultaneously to the attribution of beliefs (and in accordance to these beliefs) to whoever produces them. But these collateral beliefs typically diverge radically from those of the interpreter. That is, although we can only assign meanings relative to a body of beliefs, correct interpretations need not assimilate this background knowledge to our own. Figuratively speaking, we ascribe meanings across doxastic gaps.

Whether the result of this process is so obscure that we should review the interpretation, or whether we are just dealing with a radically different world view, is, as Putnam puts it, entirely a question of "reasonableness" (Putnam, 1988, p.75). A theory of reference would have to provide a formal reconstruction of this notion,

> and this, I have argued, would be no easier to do than to survey human nature *in toto*. the idea of actually constructing such a definition of

> synonymy or coreferentiality is totally utopian.
> **(Putnam, 1988, p.75)**

4.5 Summary

In this chapter, I argued that cognitive science lacks an explicit philosophical theory of its explanandum. That is, there is no clear account of *cognition* independent of particular (computational) cognitive models. I tried to provide insight into the relation between this problematic entanglement of empirical and conceptual issues within the discipline and its tacit *functionalist* underpinnings. Functionalism, construing mental states as (broadly) algorithmic states, turned out to be intimately related to the computational-representational paradigm underlying many theories in the field. But (machine) functionalism, as a philosophical theory of the mind, is faced with insurmountable conceptual problems.

In the remainder of this text, I want to explore the potential prospects of leaving behind functionalism and placing cognitive science on more solid philosophical foundations.

5 Cognitive Science and Inferentialism

So far, while criticizing the current state of philosophical groundwork in cognitive science, I have remained mostly implicit about the different conception of cognition that has been tacitly guiding this critique. Thus, the task at hand is to make explicit this theoretical picture and briefly explore the prospects of its application to cognitive science. Accordingly, this section for the most part is a presentation of the work of Pittsburgh philosopher Robert Brandom (Brandom, 1994, 2000), some aspects of which I intend to briefly canvass. I do not want to pretend that what follows is a comprehensive account of his philosophical oeuvre, although I hope it can serve as an introduction to his ideas, even if utterly incomplete.

5.1 Introduction

In a first attempt to locate Brandom's project within the vast realm of philosophies of the mind, one might identify it as a *normative functionalist*[1] theory of cognition, or, more specifically: of *concepts*. That is to say, concepts and the meanings associated with them are analyzed in terms of the ways in which they alter or contribute to certain *normative* facts concerning what individuals *ought* to do. Further, it can be classified as a form of *inferentialism*, giving pride of place to practices of "giving and asking for reasons" (Brandom). As I have emphasized before, there is an intimate connection between cognitive capacities and concept use. According to Brandom, although we often (correctly) attribute to nonlinguistic creatures intelligence similar to our own, such assessments ultimately are parasitic on genuinely linguistic capacities. He thus identifies as the main purpose of his project to account for the ability to master concepts, or, more precisely, to explicate what this mastery consists in. His main concern is to explain *what* counts as employing concepts, rather than *how* it is done.

The strategy pursued is a demarcational one, focusing on what is distinctive of linguistic forms of cognition (*sapience*) vis-à-vis the kind of rationality we share

[1] It may seem that, after having argued against *functionalist* theories of meaning, I simply present another such theory as a possible alternative. However, it was *machine functionalism* that was subject to criticism in the previous chapter. The distinctively normative approach followed here avails itself of entirely different conceptual resources and therefore is not directly susceptible to the kind of critique laid out before. I believe certain aspects of Brandom's conception of meaning may be incompatible with Putnam's theoretical stance, but I cannot go into detail here.

with nonlinguistic creatures (*sentience*). In a further step, an account of the latter may then be derived from the former. Initially, this order of explanation may appear unnatural. Given the fact that linguistic capacities must have arisen out of more primitive faculties in the course of human evolution, the more promising approach might seem to take the opposite direction, viz. to first explain the cognitive (or perceptual etc.) capacities common to animals and humans and then come up with a story about specifically linguistic intelligence. On this picture, sapience merely represents the topmost layer in a bundle of broadly cognitive capabilities that should be investigated in parallel. Much of contemporary cognitive science seems to adhere to similar views – recent developments in so-called "dynamic" cognitive science, in particular, bear witness to such more integrative conceptions of cognition.

I do not intend to argue directly in favor of Brandom's demarcational approach at this point. It will turn out later that rejecting an egalitarian conception of cognitive capacities does not leave us puzzled at how creatures ever managed to acquire linguistic intelligence. On the contrary, a lucid conception of sapience ultimately also offers an account of how it could arise out of more primitive abilities.

5.1.1 Sapience

What, in a nutshell, is it that is distinctive of human cognition, of *sapience*? As part of the first chapter of *Making It Explicit* (Brandom, 1994), within just two dense paragraphs, Brandom manages to outline the central ideas of his project, including several key concepts and their interrelation:

> Our transactions with other things, and with each other, in a special and characteristic sense *mean* something to us, they have a *conceptual content* for us, we *understand* them in one way rather than another. It is this demarcational strategy that underlies the classical identification of us as *reasonable* beings. ... We are the ones on whom reasons are binding, who are subject to the peculiar force of the better reason.
>
> This force is a species of *normative* force, a rational 'ought'. Being rational is being bound or constrained by these norms, being subject to the authority of reasons. Saying 'we' in this sense is placing ourselves and each other in the space of reasons, by giving and asking for reasons for our attitudes and performances. Adopting this sort of practical stance is taking or treating ourselves as subjects of cognition and action; for attitudes we adopt in response to environing stimuli count as *beliefs* just insofar as they can serve as and stand in need of reasons, and the acts we perform count as *actions* just insofar as it is proper to offer and inquire after reasons for them. Our attitudes and acts exhibit an intelligible content, a content that can be grasped or understood, by being caught up in a web of reasons, by being inferentially articulated. Understanding in this favored sense is a grasp of reasons, mastery of

> proprieties of theoretical and practical *inference*.
> **(Brandom, 1994, pp.4-5)**

Two pivotal ideas of Brandom's project can be extracted from the above paragraph:

(a) beliefs are entangled in a "web of reasons," their conceptual content is *inferentially* articulated

(b) reasons exercise a *normative* grip on us

These thoughts allow for a twofold characterization of his work as a *normative pragmatism* on the one hand, and as a *semantic inferentialism* on the other. I will now turn to a brief description of each of these aspects.

5.2 Normative Pragmatism

The force of the better reason constitutes a "rational ought" that belongs in the realm of norms, and norms are intelligible only as the product of *intersubjective* transactions. As a consequence, studying the normative structures underpinning linguistic rationality must result in a shift of scope from individuals to communities of interlocutors.

Instead of raising the question of what is required of an individual's actions in order for them to be called rational, one must look at the interactions between members of a community, in an attempt to determine criteria that settle whether their conduct can be interpreted as a "game of giving and asking for reasons" (Brandom). Again, that amounts to an inversion of the conventional order of explanation. Accounting for communicative practices is often thought to presuppose an understanding of rationality on the level of individuals. Rejecting this view, Brandom insists that the solitary exercise of concepts is parasitic on their public exchange between interlocutors. Turning the common picture on its head, what an individual does when he or she *thinks* is best understood as a kind of internalized debate.

Brandom's approach is *pragmatist* in conceiving of conceptual content as conferred on expressions by their *use*. Moreover, it is *normative* in that this use is analyzed in terms of its interaction with normative facts. More specifically, on this account, what follows from the application of a concept is an obligation (or *commitment*) whose content corresponds to the content of the concept. Grasping conceptual content consists in understanding what its application obliges one to do or say. As an example, by asserting that this patch of grass is *green*, I have thereby committed myself – among other things – to asserting that it is *colored*. Insight into this normative structure is what mastery of a concept ultimately amounts to.

But the norms governing the application of a concepts exhibit another important dimension: based on previous assertions, we may or may not be allowed (or *entitled*) to make further claims (or undertake actions). For instance, having stated that *it is raining*, I am thereby – ceteris paribus – entitled to believing that *the ground is*

wet. Clearly, collateral beliefs can interfere with this permission: maintaining that the specific area is located under a roof, for example, invalidates (again, ceteris paribus) the first claim as a license to asserting the second one.

5.3 Semantic Inferentialism

The normative relations pertaining between claims (or actions) can be interpreted as underwriting *inferential* relations. The fact that, having asserted a specific set of claims, I am either obliged or permitted to asserting certain other claims allows us to identify them as *premises* and *conclusions*, respectively.

This idea can be worked out formally by associating with each concept as its content the set of inferences it is contained in (or a subset thereof). It is this aspect of Brandom's project that allows for its classification as a *semantic inferentialism*. In privileging inference over truth (or reference) as the semantic primitive, this approach sets itself apart from the dominant paradigms in philosophical semantics.

Why is it that we should understand cognitive capacities as distinctively *inferential* capacities? Brandom invites us to consider the differences between three ways in which systems may respond to environmental stimuli. A thermostat, first of all, may reliably detect situations where the temperature sinks below a threshold value and turn on the furnace. Second, we can imagine a parrot so trained as to respond to the presence of red objects by uttering the word "red." Finally, most of us will react to this situation by correctly applying the concept "red."

These responses obviously differ in the grade of awareness involved on part of those who produce them. But what, in particular, is it that makes the third case the application of a *concept*, in contrast to the bird's reaction, which (we may imagine) could otherwise be (acoustically) indistinguishable?

> For a response to have *conceptual* content is just for it to play a role in the *inferential* game of making claims and giving and asking for reasons. To grasp or understand such a concept is to have practical mastery over the inferences it is involved in – to know, in the practical sense of being able to distinguish (a kind of know-*how*), what follows from the applicability of a concept, and what it follows from. The parrot does not treat "That's red" as incompatible with "That's green", nor as following from "That's scarlet" and entailing "That's colored." Insofar as the repeatable response is not, for the parrot, caught up in practical proprieties of inference and justification, and so of the making of further judgments, it is not a *conceptual* or *cognitive* matter at all.
> **(Brandom, 2000, p.48)**

The theory that concepts ought to be identified by their *inferential* significance is part of a bundle of interrelated ideas. If grasping the content of a concept requires (practical) mastery of a set of inferences, or conversely, if this very content is construed as constituted by proprieties of inference, another species of inference

is required besides logically *valid* inference. What is needed is a notion of *materially good* inferences, viz. inferences whose soundness depends on the contents of the concepts involved. For instance, the inference " 'Pittsburgh is to the west of Princeton' to 'Princeton is to the east of Pittsburgh' " (Brandom, 2000, p.52) is good in virtue of the content of the terms *east* and *west*: knowing that it is sound requires knowing *how* to apply these terms correctly in practice, rather than turning on specifically logical knowledge.

The conception of conceptual content and material inference, as it stands, is somewhat circular: conceptual content is cashed out as what is understood by mastering certain inferences. These inferences in turn are thought of as good by means of the contents of the concepts that appear in them. This problem is solved within Brandom's project by construing proprieties of inference as arising out of the practical *attitude* of *treating* inferences *as correct*. Eventually, this leads to the more intricate question of how *objective* norms, i.e. norms whose scope may, figuratively speaking, extend beyond our own conception of them, can nonetheless be rooted in our *subjective* attitudes towards them.[2] I shall discuss this issue later, in context with more general considerations concerning matters of supervenience.

An inference endorsed in practice can be made *explicit* in the form of a *conditional*, as a claim for which itself reasons can be given or demanded for. The tacit inferential relations corresponding to the content of a concept can thus be, at least partly, brought into the open for assessment. The conditional is paradigmatic of the *expressive* role Brandom assigns to *logical* vocabulary. Instead of thinking of logic as establishing guidelines of correct reasoning, it is conceived as providing the tools required to make the implicit norms underlying concept use accessible, as claims, within the "game of giving and asking for reasons." The explanatory role of other pieces of logical vocabulary besides the conditional can be made sense of in a straightforward fashion: for instance, negation allows us to articulate relations of material *incompatibility* between concepts. These logical locutions empower us to engage in the "Socratic" (Brandom, 2000, p.57) business of explicating and streamlining our conceptual apparatus. Expressivism about logic gives rise to yet another characterization of the conceptual: conceptual contents are those that can be made explicit and incorporated into discursive practice.

To recapitulate: Brandom's story about the nature of discursive rationality, i.e. distinctively *conceptual* awareness of the world, revolves around the concept of *inference*. What makes something a proposition, a first class citizen of the realm of cognition, is its ability to serve (possibly along other claims) as premise or conclusion in inferences. As a consequence, even beliefs that are acquired noninferentially, by perception, count as beliefs proper only insofar as they stand in inferential relations to other locutions with conceptual significance.

It is one of the chief virtues of Brandom's theory of concepts – as opposed to

[2]To be sure, the *objective* character of discursive norms, as conceived by Brandom, involves more than just attitude-transcendence. However, for the present purposes, I will focus on this particular aspect.

representational paradigms, for instance – that its explanatory primitive is not left as an "unexplained explainer." The notion of *inference* can be made sense of pragmatically, as corresponding to a specific kind of *doing*: what one does when endorsing an inference is to adhere to discursive *norms*, norms that regulate what else, given a set of claims, one is *committed* or *entitled* to believe or do. Social practices can be characterized as a "game of giving and asking for reasons" just in case participants assess their own conduct and that of others according to such proprieties of reasoning.

Among the consequences of adopting this theoretical stance is a shift of interest from *logically valid* inferences to *materially good* ones, whose soundness hinges on the content of the concepts it involves. The content that is thus implicit in discursive practice, in virtue of the participants' treating certain inferences as good as opposed to others, can be codified in conditionals, or more generally, *logical* locutions. Logical vocabulary allows us to *make explicit*, in the form of a claim, as something for which itself reasons can be asked for and given, the norms governing the use of linguistic expressions.

In this manner, the very fabric of linguistic communities is open to debate: if logical locutions are available as part of discursive practices, participant are in a position to discuss *within* the "game of giving and asking for reasons" by what rules it should be played.

5.4 Deontic Scorekeeping

Earlier I suggested that Brandom's philosophical work would establish sufficiently crisp criteria of rationality, and thus give rise to a clear-cut distinction between two issues that are constantly run together in cognitive science: on the one hand, the *conceptual* question of *what* counts as intelligent behavior; on the other hand, the broadly *empirical* question of *how* natural or artificial systems, through their causal properties, give rise to processes can be interpreted as the exercise of cognitive capacities.

So far, I have provided a rough sketch of selected themes from Brandom's philosophical work. I will now turn to a more detailed description of the species of normative practices that deserve to be called *discursive*, as conferring on expressions distinctively *conceptual* significance. Assuming the position of a (hypothetically) neutral observer studying the social practices a community is engaged in, we must ask: what kinds of doings must be discernible on part of its members that license an interpretation of these practices as a "game of giving and asking for reasons"? In consequence, what abilities must individuals master to participate, thereby exercising cognitive capacities?

5.4.1 The Nature of Norms

From the vantage point of Brandom's normative pragmatism, within discursive social practices, individuals attribute to their doings a certain normative significance. More specifically, utterances (or actions) are treated as having an effect on their *deontic status* (Brandom, 1994, p.142): what they are *committed* or *entitled* to.

To treat someone as committed or entitled to claims (or actions) is to assess their utterances and doings according to *norms*. That essentially involves a distinction between performances that are *correct* and *incorrect*, respectively, relative to these norms. As a consequence, the inferentialist story about cognition would be incomplete without a philosophical explication of the normative raw materials it draws from. If the nature of distinctively *normative* practices remains obscure, if no account is offered of what classifies behavior as correct and incorrect, respectively, the entire project is without a solid foundation. Therefore, in what follows, I will attempt to shed light on the normative underpinnings of Brandom's inferentialist theory of discursive rationality.

He extensively discusses the issue of how practices can be seen as regulated by norms in chapter 1 of (Brandom, 1994). To him, it is crucial to maintain the demarcation between individuals that act according to norms and objects whose behavior merely conforms to laws of nature (32). As a consequence, he dismisses *regularism*, which simply seeks to identify norms with *regularities* observable in practice and therefore collapses this critical distinction:

> In order to do so, it must be possible to distinguish the attitude of acknowledging *implicitly* or *in practice* the correctness of some class of performances from merely exhibiting regularities of performance by producing only those that fall within that class. Otherwise, inanimate objects will count as acknowledging the correctness of laws of physics. **(Brandom, 1994, p.32)**

But what does it take, on part of the practitioners, for them to be properly interpreted as *treating* performances as *correct* or *incorrect* in practice? What kinds of responses should one look for that qualify as classifying behavior in that way?

Brandom offers *punishment* and *reward* as candidate reactions (34-35), conceived – for instance – in terms of positive and negative reinforcement. In other words, individuals treat behavior as *correct* by responding to it in ways that are likely to increase the frequency of these performances in the future. They treat it as *incorrect* by taking actions as to reduce its future probability.

More carefully, on this picture, one should think of individuals as assessing behavior by their *disposition* to reward or punish it, in the above sense. Brandom endorses this view, albeit not in its *regularist* version – if reward and punishment are construed in terms of reinforcement, the resulting account of norms still undermines the envisaged distinction between *correct* and *incorrect* behavior. The reason behind this collapse of norms and facts is that

> assessing, sanctioning, is itself something that can be done correctly or incorrectly. If the normative status of being incorrect is to be understood in terms of the normative attitude of treating as incorrect by punishing, it seems that the identification required is not with the status of *actually* being punished but with that of *deserving* punishment, that is, being correctly punished.
> **(Brandom, 1994, p.36)**

Moreover, theories along these lines "merely put off the issue of gerrymandering" (36). That is to say, there is no fact of the matter as to what regularity is enforced by acts of reward and sanctioning, because there exist infinitely many that agree on the performances actually observed and disagree on the rest (36).

A common approach to establishing quasi-objective measures of correctness revolves around *communal* assessments of individual actions. Instead of identifying regularities in the retributive behavior of individuals, one seeks for patterns in what entire communities treat as correct or incorrect. But such theories, Brandom argues, are confronted with two serious problems (38-39). First of all, they rest on a notion of *communal assessment* that is more or less fictitious: those assessing performances are always individual *members* of the community, rather than *the* community as a whole. This objection can be addressed by picking out specific members whose status as an *authority* enables them to speak on behalf of the community. But the question of who qualifies as an authority is a *normative* matter itself. Unless the members of this privileged subset can be identified without employing normative concepts, appeals to the assessments of experts are entirely in vain. Even the status of being a community member, Brandom insists, is of a normative species. Not just anyone who conforms to certain norms can be attributed with membership, but only those that *ought to* conform. Ultimately, different incarnations of the regularist account of norms represent "attempts to bake a normative cake with nonnormative ingredients" (41).

To recapitulate: Brandom endorses a broadly retributive understanding of norms that involves several steps. First, norms implicit in a social practice are made intelligible in terms of individuals *treating as correct or incorrect* the performances of members of the respective community. Such assessments are in turn cashed in as dispositions to *reward* or *punish* individuals for their actions. In contrast to regularist theorists, however, Brandom is not committed to a naturalistic or *non*normative rendering of these notions.

Reward and *punishment*, he insists, should themselves be conceived as containing a normative component, as doing someone *good* or *bad*, respectively. One can make sense of sanctions that have an exclusively normative effect, without necessarily affecting an individual's dispositions to act in one way rather than another. In an analysis of norms, one explains some normative concepts by appealing to other, more basic ones. The hope that this process will eventually reveal *non*normative foundations is ill-conceived. It is "norms all the way down" (44). We may imagine practices in which the violation of a specific norm exclusively results in an alteration

of *normative* status, revoking certain permissions. Enforcing correct behavior, with respect to these permissions, could in turn be a matter of *nonnormative* sanctions, but this is entirely optional. One can conceive of systems of norms that only refer to yet other norms, without the possibility of a stratification whose lowermost layer can be made intelligible in naturalistic terms. In Brandom's words: sanctions can be *internal* (44) to a system of norms. It is a characteristic of such normative structures that they give rise to *holistic* relations (45). That is to say, mastering any single such norm essentially involves mastering others.

5.4.2 Norms of Deontic Scorekeeping

Brandom does not merely discuss structures of this kind to prove a point about the irreducibility of norms – it is rather that *discursive* norms, i.e. norms underlying the "game of giving and asking for reasons," are representative of this species. It is a consequence of the distinctively *inferential* significance of concepts that they must come in bundles: one cannot have one concept without having others, more specifically, those for which it can serve as premise or conclusion, respectively. This holism at the level of *semantics* matches the aforementioned holism at the level of *norms*.

I earlier tried to explain how Brandom's *normative pragmatism* lays the foundations for a *semantic inferentialism*. But while arguing that, on this picture, concepts are conceived as normative entities, by determining what interlocutors are *committed* or *entitled* to do or say, I did not elaborate on the norms that govern the "game of giving and asking for reasons" itself. Or, quite simply, I didn't offer an account of how it is played. Here is a rough sketch: The fundamental sort of move in this game is to make a claim, undertaking a *doxastic* commitment. The consequence of making a claim, of applying a concept, is its entitling or committing interlocutors to yet other claims. In order to know what moves are legitimate, players therefore need to keep track of what their fellow players are committed and entitled to, according to previous moves. These sets of commitments and entitlements, along with relations of incompatibility, can be thought of as a kind of conversational *score*. Since they correspond to the interlocutors' *deontic* statuses, the "game of giving and asking for reasons" can be conceived as *deontic scorekeeping*.

As part of these concept-mongering practices, players may challenge a fellow player's entitlement to one of her commitments p ("asking for reasons"), who must in turn vindicate her entitlement to p by either providing (further) commitments/claims fit to serve as premises for p or pointing to another player B (possibly an "expert"), whose commitment to p may equally serve as a reason for her entitlement (in that case, if the process of assessment continues, B may be called upon to vindicate entitlement to p).

The details of *how* arguments unfold *in practice* is beyond the scope of a *philosophical* project such as Brandom's. For his purposes, it is enough to provide sufficient conditions on social practices to be discernible as *discursive* practices, as conferring on expressions *conceptual* contents. On this note, what is required

for performances to count as giving and asking for *reasons* is that their normative status is assessed along two dimensions, those of *commitment* and *entitlement*, as part of a game similar to the one I just described.

Returning to the question raised above, what are the consequences of not abiding to the rules by which the game is played? What if, for example, an interlocutor A repeatedly violates the norms governing the use of a concept p, either by seeking to infer it from the wrong premises or providing it as a reason under inappropriate circumstances? As one possible consequence, conversation partners might stop treating A's claims involving p as valid moves within the game. In normative terms, A might be *precluded* from *entitlement* to using the concept p. Going further, more severe sanctions are conceivable: if, for instance, an individual constantly refuses to vindicate, upon challenge, her entitlement to claims she is committed to, interlocutors might respond by revoking her entitlement to undertake commitments altogether. Surely, within the realm of discursive norms, this amounts to the most rigorous punishment of all, effectively withdrawing (if only temporarily) from the culprit her status as a member of the linguistic community.

Again, the issue of precisely when these sanctions are appropriate, or even under what circumstances they are *in fact* imposed, is insignificant for philosophical purposes. What matters is that these punishments, first and foremost have a *normative* significance, affecting an individual's *normative*, rather than *natural* state. It is a distinctive feature of sets of discursive norms, that is, norms governing the "game of asking for reasons," that retributive actions underwriting them can be made sense of as *internal* to these systems.

5.5 Normative Attitudes and Normative Status

Exploring central themes of Brandom's project, I identified concepts by their distinctively *inferential* purport – to ascribe to mental states such as beliefs and desires conceptual contents is to treat them as standing in need of or providing reasons. On the semantic side, this guiding idea can be elaborated into a theory of contents as *inferential roles*, assigning to each concept the sets of inferences (or some privileged subset) it is involved in as its interpretant.

These inferential relations can in turn be accounted for pragmatically, as being instituted by a particular species of linguistic social practices. It is distinctive of such *discursive* practices that participants assign to their doings and sayings a certain *normative* significance, pertaining to what interlocutors *ought to* do or say. More specifically, they need to assess their conduct according to the dimensions of *commitment* and *entitlement*, i.e. with respect to what participants are *obliged* to do on the one hand and whether they have *permission* to do so on the other. Only then can their interactions be referred to as a "game of giving and asking for reasons."[3]

[3]For a detailed account of sufficient conditions, see chapter 3 of (Brandom, 1994).

Brandom's explanatory strategy of cashing in the *normative* underpinnings of this theoretical approach has been characterized as broadly *retributive*: norms are made intelligible as based on what individuals *treat* as *correct* or *incorrect*, through their dispositions to *reward* or *punish* the behavior in question. That is, at the bottom level, the theoretical entities providing the raw materials for an *inferentialist* understanding of concepts are *normative attitudes*.

However, as was argued earlier, assessing behavior is something which itself can be done correctly or incorrectly. The question of what interlocutors are *really* committed or entitled to (their *normative status*) is thus to be held apart from the issue of what they *treat* each other as committed or entitled to. But in that case, how can normative *attitudes* give rise *objective* discursive *norms* that answer to what *is*, in fact, correct or incorrect? How can concepts employed within these discursive practices represent objective features of the world, independent of the interlocutors' conceptions of their contents?

What appears to be an isolated philosophical problem is potentially devastating for attempts of utilizing inferentialist ideas in cognitive science. For according to what has been elaborated here, the attribution of original, linguistic intentionality hinges on the possibility of identifying social practices as governed by discursive *norms*. But at best, from the perspective of the theorist assessing the interactions within a community, all there is to observe are normative *attitudes*. Unless the latter can somehow give rise to normative *status*, there is no fact of the matter as to whether the social practice is correctly identified as *discursive*. Whatever *norms* are attributed by the theorist would be entirely in the eye of the beholder, imputed in what becomes an arbitrary act of interpretation.

Brandom debates this problem in connection with Daniel Dennett's notion of *stances* (Brandom, 1994, pp.55-62). According to Dennett's theory, he explains, ascribing to an individual (or any natural system) intentional states is to adopt a certain *stance* toward it. The *normative* matter of whether it is *correct* to take the intentional stance is later answered in instrumentalist terms: the interpretation is justified just in case the ascription of intentional states allows for successful predictions regarding the system's behavior. For Dennett, Brandom goes on, there is no

> distinction between actually being an intentional system and being appropriately treated as one. Intentional systems, things that have intentional states, just are whatever things it is predictively useful to adopt the intentional state toward. ... Intentional states and intentional systems are, if not in the eye of the beholder, in the successful explanatory strategies of the theorist.
> **(Brandom, 1994, p.57)**

By coupling in this way the *correctness* of adopting the intentional stance with predictive success, intentional states are purged of their perspectival aspect and become "objective" features of the (natural) world. Rejecting this particular move, Brandom is nonetheless sympathetic of the overall idea of conceiving intentional

states as instituted by *correct* intentional *interpretation*. According to this view, the ascription of intentional states is always relative to some *interpreter*, who must be in possession of such concepts as "belief," "desire" etc. That is to say, the kind of simple intentionality attributed to animals (for instance) is unintelligible without a more sophisticated species of *linguistic* intentionality. What is the distinction between these types of intentionality, and the intentional systems that exhibit them?

In some sense, Brandom argues, the intentionality displayed by simple intentional systems is only ascribed by the interpreter – the intentional significance of their behavior is entirely the result of the interpreter's activity. The conceptual contents assigned to states and performances as part of the interpretation are not intrinsic to the interpreted system.

In contrast, the intentionality attributed to linguistic communities is not derivative in this strong sense.

> If the practices attributed to the community by the theorist have the right structure, then according to that interpretation, the community member's practical attitudes institute normative statuses and confer intentional content on them; according to the interpretation, the intentional contentfulness of their states and performances is the product of their own activity, not that of the theorist interpreting that activity. Insofar as their intentionality is derivative–because the normative significance of their states is instituted by the attitudes adopted toward them–their intentionality derives from each other, not from outside the community.
> **(Brandom, 1994, p.61)**

But how, specifically, can normative *status* originate from the community members' normative *attitudes*? In a nutshell, Brandom construes this distinction as rooted in the *social* character of deontic scorekeeping. Competent interlocutors must keep two separate books: one containing the claims (commitments) endorsed by themselves, the other keeping track of commitments and entitlements they attribute to their fellow community members. Clearly, the practical significances of whether a commitment is *undertaken* by oneself or whether it is *attributed* to someone else differ profoundly. For instance, one is obliged to repair inconsistencies (*incompatibilities*) stemming from one's own commitments, whereas the ascription of incompatible beliefs to others has no such immediate practical consequences for the ascriber. Only the *undertaking* of incompatible commitments amounts to a violation of scorekeeping norms.

To summarize: following Brandom's explanatory route, what makes something an intentional system is that it is *appropriately interpreted* as an intentional system. This perspective is in danger of an infinite regress: if the contents attributed to intentional states are merely assigned by the interpreter, then how are they obtained by the interpreter herself? Brandom faces this problem by appealing to a theory of conceptual contents as *implicit* in a community's performances: meanings are conferred on expressions by the practical attitudes of those engaging in discursive

social practices, without the need for explicit stipulation. There remains a hiatus between these *normative attitudes* and *normative status*, between what participants merely *treat* as correct on the one hand, and what *is* correct on the other. But this distinction, rather than turning on an interpretation from outside the community, is *intrinsic* to the practices envisaged by Brandom – it corresponds to a difference in social perspective: competent interlocutors must discriminate between commitments *undertaken* by themselves and commitments *ascribed* to others.

5.6 Norms and Supervenience

One of the upshots of the preceding sections was that, according to Brandom, normative *proprieties* do not collapse into natural *properties*. *Regularist* theories that seek to identify norms with regularities in communal activity were thus rejected as insufficient. To put it in a slogan once more: "it is norms all the way down."

These norms were in turn construed as resulting from the practical *attitudes* of treating performances as correct or incorrect. Normative attitudes, properly structured, are sufficient to bring about objective conceptual norms. We have thus been confronted with two more or less reductive claims:

> Couched in terms of supervenience, they are the claim that settling all the facts specifiable in *non*normative vocabulary settles all the facts specifiable in *normative* vocabulary, on the one hand, and the claim that settling all the facts concerning normative attitudes settles all the facts concerning normative *statuses*, on the other.
> **(Brandom, 1994, p.47)**

Brandom's project, while being committed to the latter claim, stands in opposition to the former. That is, although normative status supervenes on normative attitudes, these attitudes cannot be given a reductive reading as physical (or functional) properties of those that exhibit them.

If an adequate account of (original) rationality has an irreducibly normative core, purely naturalistic explanations of cognition are off the table – a story about the movement of particles simply falls short of providing the desired theoretical resources. These results challenge dominant philosophical view from the cognitive scientific mainstream. In standard literature, philosophical theories of the mind are often arranged along a materialist/dualist-divide, with unconcealed preferences for those on the former side. "Dualist" theories are dismissed as promoting an untenable Cartesian dichotomy or even involving superstitious beliefs about mental phenomena. Given only these two choices, mental states *must* be physical or functional states, or so the story goes. In consequence, one is virtually "bullied" into materialism, to adopt a phrase by John Searle.

Brandom's project, with its normative underpinnings, provides no comfort to proponents of broadly materialist conceptions of cognition. Yet it is hardly susceptible to the kind of superficial criticism brought forward against the specter of

dualism. To distinguish normative proprieties from natural properties is not to distinguish *res cogitans* from *res extensa*.

In order, nonetheless, to alleviate potential concerns regarding this *normative functionalist* understanding of the mind, a few words on the relation between normative and *non*normative vocabularies will be instructive. With normative attitudes providing the building blocks of Brandom's theory, it may appear as if issues concerning natural properties were entirely out of the picture. Ascribing intentionality, it was argued, whether original or derivative, turns on *correct* intentional *interpretation*, not the identification of physical or functional properties common to intentional systems. But isn't that to say that intentionality is in the eye of the beholder after all, regardless of what has been said on the supervenience of normative status on normative attitudes?

Doubts of this kind are informed by crude misconceptions concerning the notion of *interpretation*. In this context, interpretation is regarded as imputing content on something which is otherwise devoid of meaning, as if artificially coating an object with semantics. But while sometimes interpreteds are assigned interpretants in this manner, not all cases of interpretation can be assimilated to this kind. Normally, to say that something is a matter of interpretation is not to say that it is entirely up to the interpreter – interpretations are usually constrained by criteria of adequacy that severely limit the number of choices. More specifically, it is possible to acquire beliefs concerning the appropriateness of interpreting behavior in normative terms. Not just any physical system is properly interpreted as an intentional system, even if the property of being an intentional system is not specifiable in purely physical terms.

The subject of interpretation was already touched upon in connection with philosophical theories of *computation*. As was argued there, cognitive science, questions concerning the adequacy of computational models of the mind aside, still lacks a clear understanding even of *what kind* of property it is for something to be a *computer*. While some theorists proposed relatively crisp functional properties – that turned out to be neither necessary nor sufficient – others sought to escape the dilemma by insisting that, indeed, every physical system can be interpreted as a computer. In both cases, as was argued, the envisaged distinction of computational (and thus, potentially cognitive) and noncomputational systems collapsed. But the choice between physicalist reductionism and unrestrained interpretation is neither reasonable nor compelling.

It does not follow from the fact that propositions couched in normative language cannot directly be translated into *non*normative language that the former correspond to supernatural entities from beyond the physical world. In a recent text (Brandom, 2008a), Brandom, in what he perceives as an extension of the classical analytical project, explores more complex, *pragmatically mediated* relations pertaining between sets of vocabularies. Paradigmatic of this kind are *pragmatic metavocabularies*.

> Being a pragmatic metavocabulary is the simplest species of the genus I

> want to introduce here. It is a *pragmatically mediated semantic relation* between vocabularies. It is pragmatically mediated by the practices-or-abilities that are *specified* by one of the vocabularies (which *say* what counts as *doing* that) and that *deploy* or are th *use* of the other vocabulary (what one says *by* doing that).
> **(Brandom, 2008a, p.11)**

As far as the relation between *normative* and *naturalistic* vocabularies is concerned, Brandom mentions Huw Price's "pragmatic naturalism" as pursuing a similar agenda: "He argues, in effect, that although normative vocabulary is not *reducible* to naturalistic vocabulary, it might still be possible to *say* in wholly naturalistic vocabulary what one must *do* in order to be *using* normative vocabulary" (Brandom, 2008a, p.12). A project along these lines might bridge the gap between normative and *non*normative vocabularies without collapsing it. While it may not be possible to *reduce* the former to the latter, such an analysis has the potential to provide comfort to materialists by reconciling a thoroughly *normative* account of cognition with broadly naturalistic ideas.

5.7 A New Perspective for Cognitive Science

In what follows, I intend to give a brief answer the question What is there to gain for cognitive science from adopting these views about intentionality? Although here and there I have hinted at potential dividends, nothing substantial has been offered as of yet. I want to begin with a few rather general remarks. One of the guiding themes of the present investigation was the idea (inspired by a paper by Robert Brandom (Brandom, 2008b) that I will discuss presently) that the (broadly conceptual) question of *what* counts as a distinctively *cognititive* system is to be held apart from the (broadly empirical) question of *how* the system's internal structure induces causal processes corresponding to performances of cognitive significance. This entails a criticism of much of contemporary cognitive science, where this boundary is often distorted (as was argued, this can be made sense of as stemming from a tacit *functionalist* conception of the nature of intentional states). As a consequence, cognitive theorists engage each other in debates about whether cognition should best be understood as activity in a neural network, or as algorithmic modification of syntactic items, or, recently: in terms of dynamic systems. These discussions often revolve around whether individual mechanisms are sufficiently sophisticated to reproduce certain isolated psychological phenomena, or simply how they relate to the classical computational paradigm. *Conceptual* questions about the nature of intentional states are either thought to be wholly entailed by the findings of such *empirical* research or already answered within the confines of the computational-representational understanding of mind (CRUM). But as was argued, propositional attitudes such as belief and knowledge have their place in a *space of reasons* (Sellars), instituted by *normative* justificatory relations. Ignoring this crucial lesson, cognitive science is committed to theories of the mind couched

entirely in natural scientific vocabulary. But examining the structure of brains or dissecting the perceptual apparatus of an organism is not to study *cognition* at all, at least not in a sense relevant to our understanding of *what* it is. That is not to say that, for instance, neuroscience does not constitute a legitimate area of scientific inquiry. But claims to the effect that neuroscientific facts tell us everything there is to know about intentional states must be rejected.

Cognitive science appears to be oblivious to central philosophical lessons about the mind, to the detriment of the entire discipline. In absence of a clear account of the *explanandum*, isolated phenomena are installed as ad-hoc criteria for the adequacy of models. Without knowledge of *what* cognition is, it is difficult to tell whether any specific theory is making progress in explaining *how* systems come to exhibit it.

One of the main virtues of Brandom's project, accordingly, is that it offers a lucid understanding of *what* rationality consists in, emphasizing (in line with thinkers such as Sellars and Kant) the inherently *normative* dimension of cognitive phenomena. But drawing the line between cognitive science and philosophy of mind is merely one aspect of his theory. In a paper explicitly concerned with philosophical presuppositions underlying cognitive science (Brandom, 2008b), Brandom suggests how lessons about the nature of concepts can be brought to bear on scientific research. He identifies three different levels of conceptual awareness, each of which corresponds to the mastery of a specific distinction:

(a) concepts that *describe* as opposed to concepts that merely *label*

(b) the *contents* of concepts, in contrast to the *force* of applying them

(c) *complex* predicates, to be distinguished from *simple* ones

The first point just amounts to a different way of putting the lesson that *inference* goes beyond mere *classification*, as has been argued earlier. The second roughly involves the capacity to make explicit, in the form of conditionals, the contents of concepts (allowing for hypotheticals). The third one puts interlocutors in a position to form new predicates by means of statements incorporating nested quantifiers. I am not so much interested in the details and implications of these specific distinctions, but the general idea of exploring semantic hierarchies, where the concepts located on any particular tier presuppose and elaborate abilities found on lower levels. As Brandom suggests, such structures define potential lines of research to explore for cognitive science: for instance, AI might be concerned with whether computers can be brought to exhibit these capacities; neuroscience, on the other hand, may offer theories of how the brain performs the corresponding tasks.

A fine-grained analysis of the semantics of various pieces of vocabulary and their interrelation thus directly leads to the study of cognitive systems capable of deploying these locutions, thereby displaying varying degrees of intelligence. In particular, we may raise the question of whether creatures that are merely *sentient* can be taught to adhere to discursive norms, in order to develop an understanding of how

linguistic intentionality could emerge out of less sophisticated forms of responsive awareness. It is precisely the *demarcational* characterization of *sapience* laid out by Brandom that falls into place here, putting us in a position to say what is necessary to cross the boundary.

Recent movements in the field, sometimes jointly referred to as "dynamic" cognitive science (Clark, 2001, Ch. 7 and 8), adhere to a more integrative picture of cognitive capacities. Striving for what is an inherently *anti*-representationalist and *anti*-computationalist account of intelligence, these paradigms, eschewing concepts originating in the familiar computational framework, conceive of systems as directly interacting with their environment, without the need for mediating internal representations. Such interactive processes are then modelled as dynamic systems, in terms of attractor states and differential equations, involving mathematical tools stemming from physics rather than logic.

The "dynamic" approach combines what I see as an important insight on the one hand, and an obvious shortcoming on the other. I want to argue that the latter can be remediated by paying attention to certain aspects of Brandom's theory, indicating one more way in which cognitive science may benefit from adopting an inferentialist perspective. The *insight* of "dynamic" cognitive science corresponds to the following observation: it is often possible to account for performances of simple intentional systems without appeal to the formation and computational modification of internal representations. Instead, biological mechanisms can be identified that directly exploit relevant traits of the environment. For instance, on this picture, it may turn out that the movement of limbs does not require complex algorithmic procedures, but is controlled by only a few variables within a limited space of possible movements that is highly structured by the interaction of physical parameters. It contributes much to the charm of "dynamic" cognitive science that it gets rid of certain (supposedly) hard problems in this elegant manner.

The recoil from computationalism that is manifest in this strategy is patently informed by such critics as Hubert Dreyfus (Dreyfus, 1979). On a Heideggerian reading of AI, he identifies linguistic discourse as presupposing a more fundamental kind of coping and orienting in the world, a competence whose holistic character presumably defies "rational" analysis. In the same spirit, "dynamic" cognitive science abandons talk of *representation* or *computation* in explaining the functioning of its models. And indeed, a discipline that is concerned with a natural scientific description of cognitive systems ought to jettison these concepts from its ontology. The notion of *representation*, in particular, is highly ambiguous, usually assuming a hybrid role: the term imports intentional vocabulary into theories inherently devoid of an intentional dimension. (As was argued, representations are assigned contents (or referents) to account for the conceptual contents of propositional attitudes.) However, "dynamic" paradigms, in rejecting intentional primitives, often display a reductive or even eliminative slant. That is to say, intentional interpretations are dismissed as "not acceptable" (Clark, 2001, p.129) not just as part of computational models, but in general.

And here, "dynamic" cognitive science certainly goes wrong (as critics from

within the discipline have pointed out). Surely, any theory of *cognition* ought to deliver a story about "high-level" capacities that exhibit a *representational* dimension.[4] For, according to what has been said here, in a certain and specific sense cognition just *is* concept use. Correspondingly, the task for cognitive scientific theories would be to offer an account of how systems come to demonstrate *conceptual* capacities, rather than getting rid of this explanatory burden. More specifically, the "dynamic" paradigm is indebted to showing how the direct, reciprocal forms of interaction that are its primary objects of study may give rise to abilities sufficient for engaging in discursive linguistic practices. In its current state, however, it clearly lacks the required explanatory resources – with the gap separating its models from "high-level" cognitive capacities seemingly insurmountable, reductionism becomes ever more tempting. Such is the *shortcoming* of "dynamic" cognitive science. But the commitment to reductive conclusions seems entirely optional, and more liberal interpretations may dovetail well with Brandom's normative pragmatism. Without forcing normative discourse into a naturalistic mold, "dynamic" theories have the potential to investigate into the (biological) mechanisms underlying performances in virtue of which individuals can participate in the "game of giving and asking for reasons." And it would not come as a surprise if these mechanisms were inextricably connected to other aspects of human (or animal) nature.

Brandom's project has more to offer for cognitive science than the bitter lesson that reasons and conceptual contents belong in the realm of the normative, forever out of reach of natural scientific theory. Having learned that lesson, cognitive science can move on to other, presumably more interesting issues, guided by what is a remarkably detailed philosophical account of *sapience*. As suggested, the discipline can begin to develop an understanding of how creatures (or artificial systems) can be brought to exhibit conceptual capacities, organized in semantic hierarchies. Finally, appealing to Huw Price's "pragmatic naturalism," Brandom indicates how it may be possible to *lessen the sting of denying the semantic reducibility of normative to naturalistic vocabulary by specifying in a naturalistic vocabulary what one must* do *in order to deploy various irreducibly non-naturalistic vocabularies, for example normative or intentional ones* (Brandom, 2008a, p.70).

[4]For Brandom's rendering of the *representational* dimension of conceptual content, see (Brandom, 1994, Ch. 8)

Bibliography

Brandom, Robert B. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge: Harvard University Press, 1994.

———. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge: Harvard University Press, 2000.

———. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford: Oxford University Press, 2008a.

———. "How Analytic Philosophy has Failed Cognitive Science.", 2008b. `http://www.pitt.edu/~brandom/index.html`.

Chalmers, David J. *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press, 1996.

Churchland, Patricia S., and Terrence J. Sejnowski. *The Computational Brain*. Cambridge: MIT Press, 1992.

Churchland, Paul M. *A Neurocomputational Perspective*. Cambridge: MIT Press, 1989.

Clark, Andy. *Mindware: An Introduction to the Philosophy of Cognitive Science*. New York: Oxford University Press, 2001.

Collins, Allan. "Why Cognitive Science." *Cognitive Science* 1: (1977) 1–2.

Cummins, Robert. *The Nature of Psychological Explanation*. Cambridge: MIT Press, 1983.

Cummins, Robert, and Georg Schwarz. "Connectionism, Computation, and Cognition." In *Connectionism and the Philosophy of Mind*, edited by Terence Horgan, and John Tienson, Dordrecht: Kluwer, 1991, 60–73.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge: MIT Press, 1979.

Fodor, Jerry A., and Zenon W. Pylyshyn. "Connectionism and Cognitive Architecture: A Critical Analysis." *Cognition* 28: (1988) 3–71.

Haugeland, John. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press, 1985.

Horgan, Terence, and John Tienson, editors. *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer, 1991a.

Horgan, Terence, and John Tienson. "Settling into a New Paradigm." In *Connectionism and the Philosophy of Mind*, edited by Terence Horgan, and John Tienson, Dordrecht: Kluwer, 1991b, 241–260.

McDowell, John. *Mind and World*. Cambridge: Harvard University Press, 1994.

Norvig, Peter, and Stuart J. Russell. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs: Prentice-Hall, 1995.

Poole, David, Alan Mackworth, and Randy Goebel. *Computational Intelligence: A Logical Approach*. New York: Oxford University Press, 1998.

Putnam, Hilary. *Representation and Reality*. Cambridge: MIT Press, 1988.

———. "The Content and Appeal of 'Naturalism'." In *Naturalism in Question*, edited by Mario De Caro, and David Macarthur, Cambridge: Harvard University Press, 2004, 59–70.

Searle, John R. "Is the Brain a Digital Computer?" *Proceedings and Addresses of the American Philosophical Association* 64, 3: (1990) 21–37.

Shanker, Stuart. *Wittgenstein's Remarks on the Foundations of AI*. London: Routledge, 1998.

Smolensky, Paul. "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn." In *Connectionism and the Philosophy of Mind*, edited by Terence Horgan, and John Tienson, Dordrecht: Kluwer, 1991, 281–308.

Thagard, Paul. *Mind: Introduction to Cognitive Science*. Cambridge: MIT Press, 2005.

Turing, Alan M. "Computing Machinery and Intelligence." *Mind* 54: (1950) 433–57.

Wikipedia. "AI-complete — Wikipedia, The Free Encyclopedia." `http://en.wikipedia.org/wiki/AI-complete`, 2009. [Online; accessed 27-March-2009].