How to Stop Thinking:

A Massively Modular Response to the Frame Problem

Robert Stephens

A Thesis

in the

Department of Philosophy

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Arts (Philosophy) at
Concordia University
Montreal, Quebec, Canada

August 2008

Your file Votre référence
ISBN: 978-0-494-45505-0
Our file Notre référence
ISBN: 978-0-494-45505-0

ABSTRACT

How to Stop Thinking: A Massively Modular Response to the Frame Problem

Robert Stephens

We commonly turn to the metaphor of the mind as a sort of computer, yet we are incapable of programming a computer to perform even the simplest cognitive tasks that humanity is capable of, and this stark failure speaks to the centrality of the problem of *framing*. This 'frame problem' is one of determining relevance – of limiting thought regarding an impending action to that (and only that) which falls within the context at hand – in such a way that computationally tractable thought processing can take place. The simple fact is that we *do*, in fact, do this in day to day cognition, ubiquitously and quite efficiently. Yet it is not at all clear *how* we manage to do it without entailing a constant and nearly infinite revision of the entire epistemic background, resulting in combinatorial explosion. It is a question of *how to stop thinking.*

This thesis endeavours to obviate the frame problem with a massively modular model of cognition based largely on the work of Peter Carruthers in his 2006 book *The Architecture of the Mind.* Where Carruthers' argument is vulnerable, other recent work in psycholinguistics is offered in defense and, ultimately, an account is presented explaining how we frame cognitive tasks in such as way as to adequately account for the inferential and holistic reasoning abilities we take for granted while still maintaining a materialist model that is neither strained by computational intractability, nor necessitates a central executive control mechanism, or 'ghost in the machine.'

TABLE OF CONTENTS

INTRODUCTORY REMARKS:
Framing the Discussion

> And thus the native hue of resolution
> Is sicklied 'oer with the pale cast of thought,
> And the enterprises of great pitch and moment
> With this regard their currents turn awry
> And lose the name of action (*Hamlet, III, i)*

Hamlet thinks too much. Desperately seeking a certainty of purpose and decision, he is incapable of taking any action until he has all the facts. But fact-checking can be an infinite game if one allows it to be, and in Hamlet's case, his common-sense, proto-scientific goals of empirical testing and evidentiary adjudications of certainty bring him nothing but melancholy and a quite literally *terminal* indecisiveness. There are so many questions that need to be answered and bits of evidence to be sifted before making any decision *to [act] or not to [act]. Is the ghost of his father truthful? Is his mother guilty of murder? Does Ophelia love him? Does Claudius know that Hamlet is not mad? What comes after death anyhow?* How could Hamlet ever justify an action based on a finite set of beliefs? How would he ever be certain he had deduced the correct mode of action from the relevant facts, and had not been led astray by irrelevant ones? "Nothing is either good or bad but thinking makes it so" *(II, ii)* he despairs, embracing instead a relativistic epistemology in which thought and reason are only so much shifting sand, constantly remaking the landscape. Too many unanswered questions. Too many possible implications of what he has experienced in the past, and what he intends for the future. Certainty demands that he think it all through, but circumstance cuts him short. Hamlet never succeeds in coming to a final decision, he cannot stop thinking, until his thinking is stopped and his decisions are made for him by fate and by death.

Now jump ahead roughly 400 years and, with Hamlet in mind, enjoy a fable from cognitive scientist Daniel Dennett about some frustrated artificial intelligence (AI) researchers:

> Once upon a time there was a robot, named R1 [...] Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery and precious energy supply was locked in a room with a time bomb [...] There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT (WAGON, ROOM) would result in the battery being removed from the room [...] Unfortunately, however, the bomb was also on the wagon (Dennett, 1984: 41).

The first model of R1 just goes ahead and tows out the wagon, not recognizing that removing the battery from the room *also* brings the bomb with it, thus blowing itself up. A new robot, R1D1, is developed to avoid this problem with explicit programming to consider the implications and side effects of its actions. This time, the robot does not bring the bomb out on the wagon, in fact, it does not move at all.

> It had just finished deducing that pulling the wagon out of the room would not change the color of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon – when the bomb exploded (Dennett, 1984: 42).

R1D1 had gotten hung up on irrelevant details, so the obvious answer was a redesign, R2D1, which would be programmed to ignore irrelevant implications and only act on *relevant* information.

> When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting [again], Hamlet-like, outside the room with the ticking bomb [...] 'Do something!' they yelled at it. 'I am,' it retorted, 'I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and...' the bomb went off (Dennett, 1984: 42).

Dennett's doomed robots could not reason through their existential dilemma in time to save themselves from destruction, just as Hamlet could not. They could not stop thinking and settle on a response, and both stories equally frustrate us, as we impatiently yell "Do something!" from the sidelines, thinking to ourselves how manifestly *stupid* it is to dither so much in the face of imminent danger. The dramatic irony in both stories is that we sense that *we could* judge the relevant issues and prescribe the correct mode of action to be taken, so why are these doomed protagonists unable to do the same?

The interesting philosophical question here is one of determining relevance – of limiting thought regarding an impending action to that (and only that) which falls within the context at hand – of *framing* cognitive contexts in a such a way that computationally tractable thought processing can take place. The simple fact is that we *do*, in fact, do this in day to day cognition, ubiquitously and quite efficiently. Yet it is not at all clear *how* we manage to do it. As Steven Pinker notes:

> [t]he problem escaped the notice of generations of philosophers, who were left complacent by the illusory effortlessness of their own common sense. Only when artificial intelligence researchers tried to duplicate common sense in computers, the ultimate blank slate, did the conundrum, now called 'the frame problem', come to light. Yet somehow we all solve the frame problem whenever we use our common sense (Pinker, 1997: 15).

Indeed, all philosophical work attempting to explicate or model the structure of human cognition bumps up ineluctably against the difficulty of how we determine relevance and frame cognitive contexts. In every dispute between various models of 'how the mind works', the main charge leveled – by all sides – is that opposing models cannot adequately explain how quotidian common sense reasoning can take place without entailing a constant and nearly infinite revision of the entire epistemic background, of all previously held belief, resulting in combinatorial explosion. There has to be some way for a particular thought process to 'know where to stop' – a way to frame the task, to impose frugality and avoid having to engage in exhaustive searches. Otherwise, we would all end up like Hamlet, and Dennett's R-series robots.

We do not need to focus exclusively on existential conundra, however. Pick any simple daily dilemma to illustrate the point, for example: *what shoes to wear today.* In the context of this decision process, what are the relevant things to take into account? The weather certainly; comfort; perhaps the destination (a fancy restaurant? a tobogganing adventure? a ballet recital?); the level of disrepair of the shoe; how well a certain shoe matches the rest of the ensemble; the relative merit of other shoes under consideration; whether or not the shoes belong to you… A whole host of potentially relevant details are adduced and quickly computed. Yet, as we do this, we routinely rule out a nearly infinite number of details or questions as *irrelevant* to our decision: what is the name of the person who stitched the sole? what cardinal direction is your destination? exactly how many steps will it take to get there? what is the etymology of the word 'shoe'? should you walk on all fours? Immediately, these latter questions strike us as perfectly irrelevant, and many downright silly in this context. But *how* do we know these questions are *in this case* irrelevant? (And how, exactly, do we recognize in another case that they may have suddenly become relevant?)

The simple fact is that we do, generally, know what is relevant and what is not: we tackle the relevant questions, and if satisfied, produce a decision, or judgment. Without the ability to rule out (most) questions as irrelevant, we would never be able to get anything done, as we would never be able to stop thinking and arrive at a solution. Thinking without framing would never end – it would be computationally intractable. Yet, explaining the process by which context framing takes place has been the bane of philosophers of mind and cognitive scientists. As Dennett's robot fable alludes, research into the creation of artificial intelligence has proven thus far ineffective at designing cognitive architectures capable of making relevance determinations. We commonly turn to the metaphor of the mind as a sort of computer, yet we are incapable of programming a computer to perform even the simplest cognitive tasks that humanity is capable of, and this stark failure speaks to the centrality of the problem of framing.

This *frame problem* will be the focus of this thesis. The goal will be to construct a philosophical model of how human reasoning and cognition have evolved and operate by attempting to unify empirical observations about how we generally understand (or fail to understand) context with insights from an interdisciplinary inquiry into how we psychologically and linguistically account for relevance and salience in social and communicative contexts. I will apply these accounts to the ultimate question of how we frame cognitive tasks and employ determinations of relevance in such as way as to adequately account for the inferential and holistic reasoning abilities we take for granted as exclusively characteristic of humanity, while still maintaining a thoroughgoing materialist model that is neither strained by computational intractability, nor necessitates demonlike reasoning abilities that would require a central executive control mechanism, or 'ghost in the machine.' I also intend to look into the relationship between natural language and the language of thought, or mentalese, and how context framing in natural language communication (via heuristics) may actually bootstrap conscious context framing in generalized inferential reasoning.

In chapter one, I will consider various formulations of the frame problem, focusing attention on that put forward by Jerry Fodor, who suggests that it poses a formidable obstacle to any model of cognition. Fodor argues that our distinctly human abilities of holistic abductive inferential reasoning cannot be explained until we have a clear understanding of how to solve the frame problem. The rest of chapter one will outline the turn to "massively modular" models of cognitive architecture which hold out hope of answering the frame problem, and Fodor's *a priori* objection to such models.

Chapter two will present what I believe is the most promising massively modular model of cognition in terms of rebutting Fodor's objections and settling the frame problem: that of Peter Carruthers, who proposes a model in which holistic common sense reasoning is a sort of *virtual* faculty which supervenes on a massively modular cognitive architecture. After a fairly extensive exegesis of Carruthers' account in his 2006 book *The Architecture of the Mind*, I will examine how effectively it answers Fodor, and enumerate a number of potential issues which nevertheless require further resolution.

The final chapter will take up this critical examination of Carruthers' account and attempt to resolve the issues that are problematic within it. Specifically, Carruthers' account may somewhat cavalierly abuse the role of "context" to explain how we frame relevance in such things as memory retrieval and goal-directed behaviour, which may leave his account open to accusations of circularity and question-begging. In an attempt to find a way around this potential circularity, I will focus part of this chapter on an examination of language and *pragmatics* in order to gain a deeper understanding of how we determine relevance and context *within* language and communication. I will be looking specifically at how speaker intention is ascertained in communicative contexts, and how sentential implicatures, and other 'indirect speech acts' (such as metaphor, irony and the like) are understood in conversation. As Carruthers' proposal for how we cognitively determine relevance hinges on the use of natural language, the idea here will be to salvage his account by way of language, finding a way to co-opt relevance determination heuristics from communication in natural language to apply to the workings of the language of thought.

The final chapter will also look to some recent work in psychology and linguistics in order to empirically verify the plausibility of the model set out by Carruthers, and to determine whether some predictions that could be made if his model is correct are supported by the available scientific evidence. Additionally, some evidential basis for his claims to modularity will be evaluated, including whether holistic common sense reasoning suffers from a susceptibility to some form of *systematic breakdown* that, according to Fodor, modular cognitive processes should exhibit. The ultimate conclusion will be that there is, indeed, a massively modular model of human cognition which can avoid Fodor's objections and convincingly show that the frame problem is not nearly as devastating as Fodor makes it out to be; holistic common sense inferential reasoning can be explained in a computationally tractable way without resorting to some sort of 'ghost in the machine.'

CHAPTER 1:
Frames, Frugality, and Fodor

1.1 *Framing 'the frame problem'*

Before seeking to answer the frame problem, it is best that we first define *precisely* what the problem is. Over the past 40 years, 'the frame problem' has come to mean somewhat different things, depending on the discipline in which it is being posed, and there is some debate over whether the version of it that is the focus of *this* thesis is even an accurate representation of what it was originally meant to be. Pylyshyn gives us a brief history of the problem dating back to its introduction by McCarthy and Hayes (1969), in a pivotal paper describing the problems facing artificial intelligence research:

> This problem has become known as the ***frame problem***. The name was chosen because it initially arose in connection with a particular formalism that was proposed for representing knowledge needed to reason about actions – a formalism which required statements ('axioms') that specified which properties of the world would remain unchanged when a certain action was carried out. The apparent need for such 'frame axioms' presented a serious problem because there was no limit to how many of them might be required in a reasonably complex world, and hence to the number of inferences concerning non-change that would have to be made (Pylysyhn, 1987: ix).

McCarthy and Hayes' original problem was how to determine what axioms there needed to be in a system in order to account for non-change, but 'the frame problem' has moved beyond that relatively narrow representational problem to encompass a much wider computational problem about the potential infinitude of the task. The reading of the frame problem that was laid out in the introduction to this thesis is more in accord with the latter, wider reading of the issue as a computational one. The problem faced by Dennett's fabled robots is the "where to stop thinking" problem – "Hamlet's problem" as Fodor calls it, that "if you undertake to consider a *non*-arbitrary sample of the available and relevant evidence before you opt for a belief, *you have the problem of when the evidence you have looked at is enough* (Fodor, 1987: 140).

This formulation does go farther than McCarthy and Hayes' original did, and Hayes himself says this Fodorian version "is a mistake":

> The frame problem is what is called in AI a representational problem, rather than a computational one. It is not concerned with such matters as the speed with which certain deductive searches can be undertaken, or how long it takes a robot to come to a decision, or how powerful a computer will be needed to get effective response times, or how many axioms will be needed to be stored in memory in order to get effective performance. [Even with] a magic computer which was an infinitely fast decision procedure for first order logic, so that all our computational problems were solved [...] the frame problem would still be with us [...] It is concerned with what axioms to input to the machine, not with what the machine should do with them once it has them (Hayes, 1987: 127).

Hayes, in fact, gets quite exercised over Fodor's co-opting of the frame problem, retorting that "Fodor doesn't know the frame problem from a bunch of bananas" (Hayes 1987: 132). The depth of Hayes' frustration here is a bit hard to understand, as it seems to be a fairly obvious extension of his original frame problem to make it a computational issue. Certainly, in his original formulation, the frame problem is "concerned with what axioms to input to the machine, not with what the machine should do with them", but the entire purpose of the axioms is to constrain the processing that will subsequently take place (by representing any *non-change* implicit in the task at hand), and therefore *points* towards a computational issue, at least. Dennett attempts to explain the dispute:

> McCarthy and Hayes, who coined the term, use it to refer to a particular, narrowly conceived problem about representation that arises only for certain strategies for dealing with a broader problem about real-time planning systems. Others [like Fodor] call this broader problem the frame problem [...] and this may not be mere terminological sloppiness. If 'solutions' to the narrowly conceived problem have the effect of driving a (deeper) difficulty into some other quarter of the broad problem, we might better reserve the title for this hard-to-corner difficulty (Dennett, 1987: 43).

Dennett is likely correct in his judgment here. Hayes himself admits that "one feels there should be some economical and principled way of succinctly saying what changes an action makes, without having to explicitly list all the things it doesn't change as well" (Hayes, 1987: 125). Of course, for Hayes, there *isn't* a way around having to explicitly list all of those things – it has to be done, via frame axioms. The only *problem* is determining what (and presumably how many) axioms are needed.[1]

Of course, in the real world of *human* cognition, we don't have the problem of choosing axioms, since they are already (apparently) part of our cognitive architecture, as we *already manage to intuitively understand and account for non-change.* As Dennett notes, when we perform a simple task like making a sandwich, "we know trillions of things; we know that mayonnaise doesn't dissolve knives on contact, that a slice of bread is smaller than Mt. Everest, that opening the refrigerator won't cause a nuclear holocaust in the kitchen" (Dennett, 1987: 49). We know all of these facts regarding non-change, presumably *without* an explicit list already programmed into our mental architecture. Hayes may want to keep his frame problem at the representational level of AI programming, but it seems quite natural to desire to extend it to the realm of *human* cognitive tasks, and start asking questions about how *we* manage, as thinking creatures, to process thoughts and actions without scrolling interminably through some list of mental frame axioms. This turns it into a computational problem, but one that directly flows from the same place: how to account for non-change in a dynamic situation. And that question really becomes one of determining relevance, and hence of when to *stop thinking.* Hayes argues that to extend the frame problem to this level is to really just make it part of the "Generalized AI Problem, (or *GAIP*) of "getting a machine to reason sensibly about the world"– a "hard problem" that won't have a specific answer (Hayes, 1987: 132). Again though, this is the point: the *GAIP* may be an unsolvable

[1] Ironically, a complicated computational decision in itself – one requiring a certain level of framing, in the Fodorian sense. For how can one know that one has enough, or the correct axioms? And what are the framing axioms that help *that* decision? There is the specter of a regress lurking there, which Hayes does not seem to appreciate.

problem to model and manufacture *artificially*, but if we look at it as a broader Generalized *Intelligence* Problem, (*GIP?*), we see that however hard it may be to answer, *evolution has already answered it.* We humans manage to do it. *But how?* This is the broader version of the frame problem which I will be endeavoring to answer in this thesis. Hayes' objections aside, we will now move to a fuller elucidation of what precisely I will be referring to by 'the frame problem,' using Fodor as a guide.

Fodor brushes aside Hayes' criticisms, noting that,

> the frame problem is so ubiquitous, so polymorphous, and so intimately connected with every aspect of the attempt to understand rational nondemonstrative inference, that it is quite possible for a practitioner to fail to notice when it is indeed the frame problem that he is working on [...] Which would be OK except that if you are unable to recognize the frame problem when as a matter of fact you are having it, you may suppose that you have solved the frame problem when as a matter of fact you are begging it (Fodor, 1987: 142).

Case in point, Fodor looks to the problem of *updating* belief in a changing world, which is ostensibly what Hayes' original formulation of the frame problem was concerned with axiomatizing. Fodor points out that most AI research works with a "sleeping dog" strategy that explicitly rules everything *unchanged* that is not directly changed as the result of action (I.e. the vast epistemic background is treated as a sleeping dog, and we let it lie there, undisturbed). As Fodor notes, "You can rely on metaphysical inertia to carry most of the facts along from one event to the next" (Fodor, 1987: 142). Yet this hardly seems satisfactory, because the sleeping dog strategy would only work if one could somehow determine objectively which beliefs remain unchanged, and assign them the status of sleeping dogs. Of course, this process has its own computational load, which will negate the effort saved by ignoring those beliefs once they are tagged as unchanged. Fodor goes further to suggest that even if you *could* identify the sleeping dogs, there are still potentially infinite "kooky facts" that could be part of the changeable epistemic background, and therefore part of the calculation as to what remains unchanged through time. He proposes a speculative property of physical particles he calls being a "fridgeon":

> I define 'x is a fridgeon at t' as follows: *x is a fridgeon at t iff x is a particle at t and my fridge is on at time t.* It is, of course, a consequence of this definition that, when I turn my fridge on, I CHANGE THE STATE OF EVERY PHYSICAL PARTICLE IN THE UNIVERSE; namely, every physical particle becomes a fridgeon [...] I repeat the moral: Once you let representations of the kooky properties into the database, a strategy which says 'look just at the facts that change' will buy you nothing; it will commit you to looking at indefinitely many facts (Fodor, 1987: 144).[2]

Again, the point to stress is not that it is impossible to update belief in a computationally expedient fashion – we do, in fact, do this all the time – the question is whether this framing is somehow axiomatic, and whether there is a way to model it (and, by extension, potentially recreate it in AI).

Belief revision, or updating, is a major component of the frame problem, as couched in Fodorian terms, even if you *don't* include the so-called "kooky facts" in your database. Imagine a person (or a thinking machine) with 150 'beliefs' – 150 factual propositions in its database. Now imagine that person is presented with a novel proposition – a new fact – and has to decide whether to believe or disbelieve it, and incorporate or reject the belief from the database accordingly. The logical assumption would be that the belief needs to be checked against the current epistemic background, in this case the 150 beliefs already in the database.

[2] There is an interesting side argument to make here that this whole discussion of a "kooky fact", like the particle property of being a fridgeon, is at odds with Fodor's writing elsewhere that pushes back quite strenuously on "unobservables" and insists on the primacy of observable phenomena in any scientific theory or discourse. In "Observation Reconsidered," Fodor concludes that "belief in the best science is rational because it is objective, and it is objective because the predictions of our best theories *can be observed to be true*" (Fodor, 1984: 42).

Now, if Fodor believes this, then it seems as if any factual database that needs updating in cognition should *not* include "kooky facts", as they are by definition *not* objectively observed or observable. So, on a computational level, one's epistemic background checks in decision making need not be *indefinite* – they should only have to survey the (relevant) background of objective observable facts that are upheld by theory. This will be a *finite* list, which is not to say that it will still be computationally feasible to perform the operation. It just seems odd that Fodor would resort to kooky facts to prove his point, when he thinks kooky facts are not even facts, and shouldn't be a part of legitimate scientific discourse.

But a logical truth-table analysis of 151 propositions would take 2^{151} lines. That's a great deal of checking. And even if each line could be computed in, say, 1/100th of a second (which seems unreasonably fast), such a cursory consistency check would take roughly 9×10^{35} years to complete (just under a billion billion billion billion years). Of course, most humans over the age of two probably have a lot more than 150 propositional beliefs to keep track of. It's easily apparent that we certainly *don't* perform this kind of consistency check when we engage in belief revision.

We need, as Clark Gylmour suggests, to formalize some type of "Computability Constraint" if we are to explain how human minds (and the possible AI machines of the future) succeed where classical propositional logic obviously fails to solve the frame problem in a feasible fashion. "The Computational Constraint requires novel and unobvious solutions, and makes everything harder and perhaps more fun" (Gylmour, 1987: 75). In the past twenty years, there has been a great deal written in cognitive science looking for that elusive novel and unobvious solution. Whether *fun* or not, there is little dispute that it has proven a *hard* problem to crack. As we have seen, it is a hard problem to even *define*, given that those who coined the term are no fans of how it has been co-opted, and the scope of the problem and availability of possible avenues of solution vary according to which formulation is being employed. In the following section, I will spell out the specifically *Fodorian* version of the frame problem which will form the basis of the inquiry in this thesis.

1.2 *Fodor's version of the frame problem*

Fodor has a particular fondness for the frame problem, as he views it to be one of the most criminally neglected and overlooked problems facing overzealous cognitive scientists. The title of Fodor's 2000 book, *The Mind Doesn't Work that Way,* is a riposte to

Steven Pinker's *How the Mind Works*, which Fodor evidently views as arrogantly and incorrectly titled, fundamentally ignorant of the frame problem, which, as quoted in the previous section, Fodor argues is "so ubiquitous, so polymorphous, and so intimately connected with every aspect of the attempt to understand rational nondemonstrative inference" (Fodor, 1987: 42).[3] Fodor gets quite exercised about what he calls the "New Synthesis" school of cognitive science, as typified by Pinker and Henry Plotkin, who "combine computational theory of mind [CTM] with a comprehensive psychological nativism and with biological principles borrowed from a neo-Darwinist account of evolution" (Fodor, 2000: 2). He believes this takes the computational model too far afield from what we actually ***know*** about how the mind works and what is ***plausible*** about the way our cognitive architecture is wired. He explains:

> Over the years I've written a number of books in praise of the Computational Theory of Mind. It is, in my view, by far the best theory of cognition that we've got; indeed, the only one we've got that's worth the bother of a serious discussion. There are facts about the mind that it accounts for and that we would be utterly at a loss to explain without it; and its central idea – that intentional processes are syntactic operations defined on mental representations – is strikingly elegant. There is, in short, every reason to suppose that the Computational Theory is part of the truth about cognition.
>
> But it hadn't occurred to me that anyone could think it's a very *large* part of the truth; still less that it's within miles of being the whole story about how the mind works (Fodor, 2000: 1).

Fodor argues that there is a "large crack in the foundations of New Synthesis cognitive architecture" that much current discourse in cognitive science seems blithely unconcerned with: namely, the idea that "maybe the computational theory of mental processes doesn't work for abductive inferences" (Fodor, 2000: 41). The objection boils down to a rather simple point: much of our day to day cognition appears to rely on

[3] Fodor chides Pinker for not even having "the frame problem" in the index to his book, which turns out not to be true, as Pinker replies in his follow-up to Fodor, "So How *Does* the Mind Work?" (2005). Fodor concedes that he was wrong, and that there are indeed two mentions of "the frame problem" in Pinker's book, but notes wryly that *Star Trek* is listed in Pinker's index *seven* times (Fodor, 2006).

abduction – utilizing global processes to make holistic rational inferences, inferences "to the best explanation," when multiple variables and courses of action present themselves. However, according to what Fodor terms the Classical model of CTM, all mental processes operate *locally*, and the type of *global* process that abduction implies seems simply impossible if the CTM is correct and complete. Fodor argues this a "terrible problem for cognitive science" (Fodor, 2000: 41) as it leaves

> the question of how to reconcile a local notion of mental computation with the apparent holism of rational inference; in particular, with the fact that information that is relevant to the optimal solution of an abductive problem can, in principle, come from anywhere in the network of one's prior epistemic commitments (Fodor, 2000: 42).

Here we have a clear instance of the type of frame problem Fodor wants cognitive scientists to answer – one that does not rely on the positing of "kooky facts" and is not even an epistemological argument, but is, rather, based on an observed phenomenon of human reasoning: holistic, abductive inference. Inference to the best explanation implies an ability to know *where to stop thinking* – to survey the epistemic background, and be able to determine the relevant information to bring to bear on the calculation; to disregard the irrelevant data, and even to *weigh* the relative relevance of data in order to find the *best* explanation; (not to mention, to be able to somehow justify quasi-objectively the quality of the explanation in order to assign it a value judgment of 'best'). On a computational level, this appears to be a completely intractable task without the presence of some sort of central executive function that is capable of such epistemic oversight and judgment – some sort of ghost in the machine. Fodor is not arguing that there *is* such a ghost, but he is suggesting that cognitive science, in particular the "new synthesis" school of computational cognitive science, is fully haunted by the specter of abduction and has offered no plausible way to account for the framing that goes on in holistic reasoning. Not only has cognitive science failed to answer this frame problem so far, according to Fodor, the immediate *prospects* of solving it look exceedingly bleak.

> I'm quite prepared to admit that it may yet turn out that all cognitive processes reduce to local ones, and hence that abductive inference is after all achieved in some way that Classical computational psychology can accommodate. But nothing of the sort is currently on offer, and I wouldn't advise you holding your breath (Fodor, 2000: 46).

Fodor concludes that cognitive science is at an "impasse" and that one would be best to "concentrate one's research efforts in those areas of cognitive processing where the effects of globality are minimal" (Fodor, 2000: 52-53). In the next section, we will look at one such area -- modular processing – as it may hold a key to answering Fodor's frame problem regarding holistic reasoning.

1.3 *(Massive) Modularity as an answer to the frame problem?*

This section will look specifically into the suggestion that a *massively modular* model of cognitive architecture could settle the frame problem and explain how day-to-day cognition, decision-making and common sense reasoning might take place while remaining computationally tractable. However, before delving deeper into the question of how massive modularity may help solve the frame problem, we must first take some time to outline what exactly mental *modules* are.

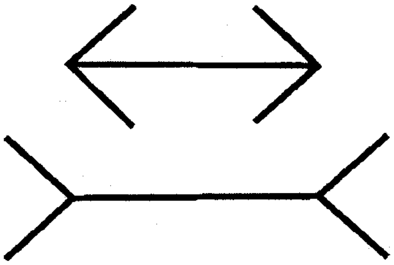
Fodor notes that he may have muddied up the definition of 'module' by lifting the term from Chomsky, but using it in a slightly different way. For Chomsky, a *module* is simply an innate information database (such as grammatical rules), whereas Fodor widens the term to encompass the dedicated processing mechanism *attached* to a particular Chomksian module. For Fodor, a module is a functional mechanism, but not just *any* functionally individuated mechanism; a Fodor-module is qualified by very specific conditions.[4] Fodor lists five questions which must be answered to determine whether a cognitive system can be considered modular:

1. Is it domain specific, or do its operations cross content domains?
2. Is the computational system innately specified, or is its structure formed by some sort of learning process?
3. Is the computational system 'assembled' (in the sense of having been put together from some stock of more elementary subprocesses) or does its virtual architecture map relatively directly onto its neural implementation?
4. Is it hardwired (in the sense of being associated with specific, localized, and elaborately structured neural systems)?
5. Is it computationally autonomous, or does it share horizontal resources (of memory, attention, or whatever) with other cognitive systems? (Fodor, 1983: 36-37)

In answer, Fodor states that "modular cognitive systems are domain specific, innately specified, hardwired, autonomous, and not assembled" (Fodor, 1983: 37). Furthermore, any cognitive systems that meets these five criteria may be said to be *informationally encapsulated* insofar as it draws inputs only from a specified domain, and its processing is impenetrable to the rest of the mind; no information outside of the specified domain can be accessed by the processor or supervene on the process. The input-output system essentially forms a computational 'black box' where feedback and exchange with other cognitive mechanisms is cut off. Sensory input systems, such as vision and phonological parsing, are held up by Fodor as clear examples of such encapsulated modular mechanisms. Visual input is quite clearly encapsulated, a fact which can be demonstrated by the characteristic patterns of *breakdown* visual perception is susceptible to – what we commonly refer to as optical illusions.

[4] Fodor makes a great deal of the distinction, and is careful *not* to use the term 'module' to apply to any and every functionally individuated mechanism in the mind. This is the mistake he believes many others have made in his wake, widening his term to the point where it no longer applies (Fodor, 2000: 56). This is an issue which will come up repeatedly throughout this thesis.

The Muller-Lyre illusion is a prime example:



Despite the lines being the same length, our visual system is tricked into seeing the bottom line as longer, as the arrows trigger an edge-detection function in our 3-D visual perceptual apparatus which suggests the bottom line is *farther away* and hence must be *longer*, once the distance effect is calculated. The interesting facet of this illusion with regard to encapsulation is the fact that the illusion *persists* even when it is *known* to be illusory. As Fodor notes, "the very same subject who can tell you that the Muller-Lyre arrows are identical in length, who indeed has seen them measured, still finds one looking longer than the other" (Fodor, 1983: 66). Background knowledge that the lines are identical fails to penetrate visual perception to correct the illusion. As a result, one can *know* the lines are identical yet still fail to *see* that they are – the visual system keeps stubbornly outputting the perception that they are dissimilar. For Fodor, this proves that "perceptual processes are 'synchronically' impenetrable by – insensitive to – much of the perceiver's background knowledge. Your current sophistication about the Muller-Lyre is inaccessible to the module that mediates visual form perception and does not, therefore, dispel the illusion" (Fodor, 1984: 39). This is the essence of informational encapsulation: the processor is to a certain extent *stupid* – the processing is entirely local and the output cannot be amended by bringing any additional information to bear – and this stupidity is what makes encapsulated processors so susceptible to characteristic patterns of breakdown or misfire.

Encapsulation does have an *upside* for cognitive processing, however. "Presumably," offers Fodor, "what encapsulation buys is speed [...] at the price of unintelligence" (Fodor, 1983: 80). The encapsulated module is constrained from getting sidetracked with processing information from other parts of the mind, which could cause it to bog down.

> In a nutshell [...] the more encapsulated the informational resources to which a computational mechanism has access, the less the character of its operations is sensitive to global properties of belief systems [...] Nothing affects the course of computations of an encapsulated processor except what gets inside the capsule; and the more the processor is encapsulated, the less information that is. The extreme case is [...] the reflex; it's encapsulated with respect to all information except what's in the current input. So it operates entirely without computing (Fodor, 2000: 63-64).

Indeed, Fodor analogizes encapsulated perceptual modules to reflexes, in the sense that modules generally mimic reflexes insofar as being *mandatory* and *fast*. As in the case of the Muller-Lyre illusion, we see the lines as dissimilar *reflexively*, and are just as incapable to overruling the perception with other globally available background information as we are incapable of overruling reflexive blinking by mental willpower.

It should be clear then why modules are relatively immune to the frame problem as Fodor has formulated it, as "to the extent that the information accessible to a device is architecturally constrained to a proprietary database, it won't have a frame problem and it won't have a relevance problem" (Fodor, 2000: 63). An *unencapsulated* mechanism, with "unconstrained access to the cognitive background" (Fodor, 1994: 216) would be hopelessly bogged down and begging the frame problem, whereas the encapsulated module has an innately specified frame which allows for reflexively fast, tractable computation.

However, perceptual systems are not what 'the frame problem' generally has in mind: we *can* build robots that can mimic human visual perception, for example – this isn't even really artificial *intelligence*, as the Fodorian sketch of visual perception made above treats that visual system as fundamentally *unintelligent*. The frame problem is, rather, one that bedevils global holistic reasoning, specifically *not* perceptual or reflex systems, so *modularity* can only serve as an answer if it is assumed *that the entire mind is modular in function.* The thesis of 'massive modularity' *(MM)*, suggests exactly this: that the mind is essentially an agglomeration of myriad modular cognitive mechanisms working in concert, yet for the most part independently from one another in terms of processing. If all cognitive processes could be shown to be modular, this could allow for a way to maintain computational tractability in all aspects of cognition.

Although he will eventually disavow it, Fodor initially entertains the prospect that massive modularity *could* account for the seemingly global abductive reasoning that takes place in day-to-day cognition. As we have seen, since modules are informationally encapsulated and domain specific, they don't have a frame problem – they work on proprietary and delimited databases (*Chomskian modules)* that solve the problem of 'where to stop' – frugality is maintained, as exhaustive search would be unnecessary. The module is local in this sense, and need not appeal to any global processing; abduction is simply banished, "because, in point of architecture, only what's in its database can be in the frame [...] it doesn't have to treat framing as a *computational* problem" (Fodor, 2000: 64). On the massively modular model, there is no 'choosing' between modules of a sort that requires abduction – there are simply *a lot* of modules, each dedicated to solving a particular problem, and the seeming 'inference to the best explanation' is nothing more than the automatic triggering of the appropriate module by the inputs it receives. "[T]hat there is nothing in the mind that can ask questions about which solution to a problem is 'best overall,' that is, best in light of a creature's beliefs and utilities" (Fodor, 2000: 64).

However, Fodor thinks the massive modularity thesis *(MM)* is not going to work as an answer:

> I'm going to argue that there's no a priori reason why MM *should* be true; that the most extreme versions of MM simply *can't* be true; and that there is, in fact, no convincing evidence that anything of the sort *is* true. In sum, no cheers for MM (Fodor, 2000: 64-65).

Fodor believes that only our peripheral, perceptual systems are modular, and that most of our actual *thinking* has to be effectuated by *non-modular* domain general mechanisms, insofar as holistic thinking appears to be unencapsulated and have unconstrained access to the global epistemic background. Indeed, he suggests the frame problem proves this fact to us:

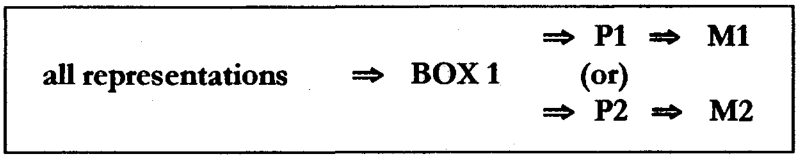
> when we try to build a really SMART machine – not a machine that will parse sentences or play chess, but, say, one that will make breakfast without burning down the house – we get the frame problem straight off. This, I argued in *Modularity of Mind,* is precisely BECAUSE smart processes aren't modular [...] In short, that the frame problem breaks out here and there *but does not break out everywhere* is itself an argument for differences in kind among cognitive mechanisms. We can understand the distribution of outbreaks of the frame problem *on the hypothesis* that it is the chronic infirmity of rational – hence, unencapsulated; hence *non*-modular – cognitive systems (Fodor, 1987: 141-142)

Massive modularity, "the idea that modularity is the *general* case; that *all* cognitive processing is informationally encapsulated," is simply implausible for Fodor, who calls it "modularity theory gone mad" (Fodor, 1987: 141). In his 2000 book, Fodor goes further and sets out a powerful *a priori* argument against a massively modular cognitive architecture -- the so-called "input problem" – which will be the focus of the next section.

1.4 *The Input Problem*

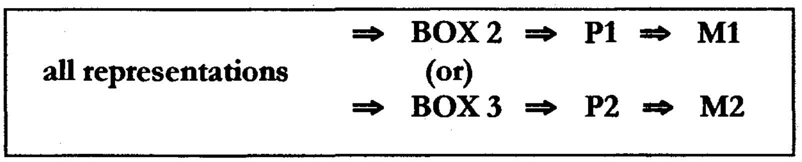
To illustrate what he sees as a fundamental flaw in the massive modularity argument, Fodor asks us to imagine a simple set up with two encapsulated modules, *M1* and *M2,* that act on representations *P1* and *P2* respectively. *M1* "turns on when and only when it encounters a *P1* representation and *M2* turns on when and only when it encounters a *P2* representation. We therefore infer that *P1* and *P2* are somehow assigned to representations prior to the activation of *M1* and *M2*" (Fodor, 2000: 72). Fodor then asks a simple, but potentially devastating question: "Is the *procedure that effects this assignment itself domain specific?*" (Fodor, 2000: 72). The following diagram illustrates the question and the two ways it could be answered:

1.



but then, this BOX 1 which assigns P1 and P2 cannot be modular...

2.



but this one courts a regress; how do representations get assigned to *BOX 2* or *3*?

In order to assign representations to type *P1* or *P2*, thereby framing the problem and routing them to the appropriate module for processing, we must postulate some *BOX* that does the sorting and assigning. Fodor's point is that the *BOX* must necessarily be *less* modular and *less* domain specific than the modules it is sorting representations for. It appears to spark a vicious regress, as, ultimately, it seems as if you are always going to need some kind of domain general *BOX1* which can take in *all representations* and begin the assignation process. Fodor concludes that "each modular computational mechanism presupposes computational mechanisms less modular than itself, so there's a sense in which the idea of a *massively* modular architecture is self-defeating" (Fodor, 2000: 73). The only way around this input problem, he suggests, would be to argue that "it's the *sensory* mechanisms that block the regress. In effect, your sensorium is assumed to be less modular (less domain specific) than *anything else in your head*" (Fodor, 2000: 74). This empiricist solution, however, is probably not one that most modularity fans, or nativists of any stripe, would be content with.

Fodor entertains what an immediate objection may be to his input problem: namely, that it flies in the face of experimental data which *seems* to show modularized, encapsulated reasoning. The data in question deals with the Wason selection task performance effects as analyzed by Cosmides and Tooby, in which people reason *better* in some (social) situations than others (Cosmides, 1989; Tooby, Cosmides, Barrett, 2005).[5]

The Wason selection data was puzzling when first uncovered experimentally in the late sixties, as it "demonstrated that reasoning performance on distinct tasks that require the use of a single rule of deductive inference varied as a function of the content plugged into the inference rule" (Clarke, 2004: 8). In the experiment, subjects were presented with cards, each with a letter on one side and number on the other, and were given the following rule:

- ***If a card has a vowel on one side, then it has an even number on the other side***

Subjects were then shown the following four cards, and asked to determine which (and only which) cards needed to be turned over to check if the rule was being followed:

- *E K 4 7*

Most subjects recognize the need to turn over the card with the vowel, but many fail to logically determine the need to turn over the odd-numbered card. Indeed, if the odd-numbered card turns out to have a vowel on the flip side, then the rule will have been violated.

So far, one might simply say it's a tricky logic puzzle. But the interesting data comes when the abstract logical relationships between symbols and the given rule is substituted by

[5] A truly complete discussion of the Wason selection task data would be beyond the scope of this chapter, but there are a number of treatments on how that data fits into this discussion to be found, such as Cosmides, 1989; Cheng & Holyoak, 1989; Clarke, 2004. The original publication of the experimental results can be found in Wason, 1968.

more meaningful, concrete items. In the second run, subjects are given cards with the names of English cities on one side, and forms of transportation on the other, along with the rule that:

- ***Trips to Manchester must be made by train***

Subjects are then presented with the cards:

- ***Manchester   Sheffield   Train   Car***

In this case, again, most subjects recognize the need to turn over the ***Manchester*** card, as that card should say ***Train*** on the flip side, if the rule is being followed. What is interesting, however, is that the content on this second set of cards elicits *much* better performance in terms of turning over the ***Car*** card in order to check that it did *not* say ***Manchester*** (which it should not, according to the rule). These 'content effects' were initially quite puzzling, as the logical form in both versions of the experiment is identical, and yet the success in deducing the correct answer to the problem varies significantly based on the specific content plugged into the conditional rule.

> This violates the most fundamental idea of formal logic, namely, that arguments are valid purely as a function of their abstract form regardless of their content. That humans consistently fail to observe the content-neutrality aspect on deductive reasoning tasks came as an enormous surprise [...] Realistic or familiar materials produce much better results than abstract or unfamiliar materials, regardless of the fact that distinct experiments employed generalizations with the same logical form and truth conditions (Clarke, 2004: 9).

Clarke goes on to explain that although the content effects were originally explained by Wason and Johnson-Laird (1983) as being a result of familiarity with concrete terms (as opposed to abstract symbols), Cosmides and Tooby re-evaluated the data and suggest instead that the content effects are a result of the presence of a "social contract" in the selection task (Clarke, 2004: 9). Cosmides and Tooby propose that natural selection has hard-wired an ability to reason more accurately and acutely in situations of social exchange, especially when the possibility of being cheated is present.[6] They also suggest that this social exchange reasoning capacity is likely modular and encapsulated, which would explain why it functions so efficiently, but the deductive successes it brings do not readily transfer to other, non-social milieu (or to abstract logical reasoning). Indeed, such a mechanism, or "cheater detection module" (*CDM*) would appear to fit the description of an encapsulated module, insofar as it is domain-specific, seems to operate subdoxastically, and its algorithm is not generalizable for use in other logically equivalent situations.

Now, to return to the main issue of this section, remember that the Wason selection data was raised by Fodor as a possible line of objection to his claim that the input problem defeats modular reasoning mechanisms. If Fodor's input problem *is* a serious challenge to massive modularity, how can he explain these performance effects on the Wason selection task? In response, Fodor takes aim at Cosmides and Tooby's arguments concerning the ostensibly encapsulated CDM, and he inverts their argument to propose that such a CDM is a perfect illustration of the input problem at work, rather than an argument against domain generality.

Fodor sets up the CDM argument briefly:

> [O]ne of the things that's supposed to make the CDM modular is that it normally operates only in situations that are (taken to be) social exchanges. Its operation is thus said to invoke inferential capacities that are not available to the mind when it is thinking about situations that it does not take to be social exchanges [...] So then, then CDM computes over mental objects that are marked as social exchange representations, and its function is to sort them into distinct piles, some which represent social exchanges in which cheating is going on, and others which do not (Fodor, 2000: 75).

[6] The evolutionary advantage of such an ability should be evident, as those who could navigate the treacherous negotiations of pre-historic human interaction without getting cheated, bamboozled or otherwise taken advantage of would be more likely to survive in the hard-scrabble hostile world of the Pleistocene.

Fodor then brings the input problem to bear on this, asking *how* representations get tagged as "social exchanges" in order to be routed to the CDM – and whether the mechanism that does this sorting and tagging is *itself* modular, though obviously in some way *less* domain specific than the CDM it routes to.

> Figuring out whether something is a social exchange and, if it is, whether it's the kind of social exchange in which cheating can be an issue (not all of them are, of course) involves the detection of what behaviorists used to call Very Subtle Clues. Which is to say that nobody has *any idea* what kind of cerebration is required for figuring out which distal stimulations are social exchanges, or what kinds of concepts that kind of cerebration would need to have access to. [...] So the massive modularity thesis can't be true unless there is, inter alia, a module that detects the relevant Very Subtle Clues and infers from them that a social exchange is going on. [...] figuring out whether something is a social exchange [...] takes *thinking*. Indeed, it takes the kind of abductive reasoning that, by definition, modules don't do and that Classical computations have no way to model (Fodor, 2000: 76).

Fodor looks at language modules as a further example. He notes that we still don't have a full understanding of how language modules (which are likely modular, in his view) *receive* the appropriate input. It is assumed that there are psycholinguistic telltales that the sensorium can detect and tag as "language" – but even this is an incomplete understanding, and doesn't begin to explain how we account for things like sign language or reading (Fodor, 2000: 77). Fodor's point in bringing up language is that

> [I]t is *much* more plausible that you don't need to do any complicated thinking to decide that an input belongs to the language domain than that you don't need to do any to detect inputs in the domain of the CDM [...] because language perception [...] can be detected psychophysically [...] and yet it turns out that empirical solutions of the input analysis problem aren't easy to come by *even* in the case of likely candidates like language (Fodor, 2000: 78).[7]

And by extension, it is totally implausible that there are simple tagging explanations in the much more complex operations of the CDM. Fodor concludes that "massive modularity is a coherent account [...] only if the input problem [...] can be solved by inferences that aren't

[7] A great deal more discussion of the language perception faculty and the question of its modularity will follow in chapters two and three.

abductive (or otherwise holistic); that is by domain specific mechanisms. There isn't however, any reason to think it can" (Fodor, 2000: 75).

Fodor asserts that all the models of computational or massively modular mental architecture that he has looked at suffer from "terminal abduction":

> So, if it's right that the New Synthesis requires the Classical model of computation, and if it's right that the Classical model of computation works only for local computations, and if it's right that only modularized processing is likely to be local in the relevant respects, then you probably can't save the New Synthesis by assuming that cognitive architecture is massively modular. By all the signs, the cognitive mind is up to its ghostly ears in abduction. And we do not know how abduction works. So we do not know how the cognitive mind works; all we know anything much about is modules (Fodor, 2000: 75).

This presentation of abduction as a serious roadblock to any account of how a massive modularity theory of cognitive architecture could answer the frame problem seems to be quite solid: any plausible massively modular model will need to tell a coherent story of how (apparently) abductive inference can work in a modular system. If there is an account that *can* be shown to explain how abductive inference works within a modular system, then the frame problem would seemingly be a long way towards being settled. In the next chapter, we will turn to one such account to see how it fares.

CHAPTER 2:
Modular Abduction

The central focus of this chapter will be to examine in depth one massively modular model of cognitive architecture which I will argue is quite effective at avoiding Fodor's input problem and establishing how a modular mind could account for the appearance of holistic abductive reasoning in a computationally tractable way. The model that will be examined is that of Peter Carruthers as set out in his 2006 book *The Architecture of the Mind.* The first section of this chapter will consist of a fairly in depth exegesis of Carruthers' account in which he posits "a sort of *virtual* faculty of inference to the best explanation [that] can be constructed from principles involved in the assessment of linguistic testimony and the interpretation of speech" (Carruthers, 2006: 386). The second section of this chapter will pit Carruthers' model against Fodor's concerns in order to show where Carruthers is effective in allaying Fodorian objections regarding the possibility of abductive inference in a massively modular mind, and whether Carruthers' account can answer Fodor's 'input problem'. At the end of the chapter, I will enumerate a number of potential problems with Carruthers' account, and point to one possible element of circularity in his treatment of relevance determination, which will be the focus of inquiry in chapter three. I will also note some aspects of his model which require empirical verification and/or a review of experimental evidence to determine if they are plausible, a task which will be also be taken up in the final chapter.

2.1 Carruthers' Architecture

In *The Architecture of the Mind*, Peter Carruthers builds a careful and detailed story of the development of a massively modular cognitive architecture designed by selective pressures in such a way as to explain the "distinctively human" aspects of human thought such as practical reasoning, creativity, abductive inference, and scientific theorizing. In the end, he disclaims that his account "is somewhat comparable to the suggestion that the formation of hurricanes probably has something to do with local increases in ocean surface temperatures – true and fruitful, perhaps, but hardly deserving of being called an explanation itself" (Carruthers, 2006: 331). He does however conclude, directly *contra* Fodor, that

> since there are powerful arguments supporting the massive modularity hypothesis, and since there are no convincing arguments against it (in particular, since the 'How possibly?' challenge can be answered, in outline), a massive modularity theorist is what every good cognitive scientist should be [...] For that is where the future action, and future success, is likely to be (Carruthers, 2006: 417).

For Carruthers, the story of the evolution of human cognition requires that it consist of interconnected modularized processors, much akin to biological systems, with specific functions and domains "built by co-opting and connecting in novel ways resources that were antecedently available in the service of other functions" (Carruthers, 2006: 21). Before even beginning to sketch his own constructive argument, Carruthers notes that the alternative to an MM model of the mind – one which posits some central 'global learning' device in place of a plenitude of adapted modules – is, aside from being evolutionarily unlikely, also potentially *disastrous*, insofar as any malfunction or injury to this device would completely cripple the mind (Carruthers, 2006: 27). On the contrary, there is much evidence from research involving brain injury which shows that this is not the case: there are innumerable ways in which *some* cognitive functions can be impaired while leaving others intact. From psychology text staples like poor Phineas Gage with an iron rod punched through his frontal lobe, to more recent studies of the very specific deficits which accompany localized injury to either Wernicke's or Broca's area of language specialization, examples of this abound.

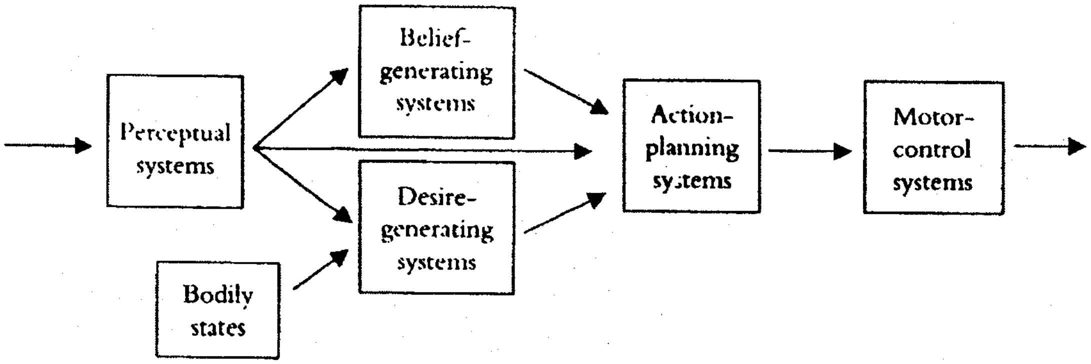
Carruthers begins his argument for massive modularity by looking first at the basic cognitive structures we inherited from our animal ancestors, both direct and distant, which demonstrate "that perception / belief / desire / planning / motor-control architectures are of very ancient ancestry indeed, being present even in insects and spiders" (Carruthers, 2006: 65). Much insect life operates on fixed-action schemata, which serve as de facto encapsulated cognitive modules, and have been empirically documented in numerous studies, such as Gallistel's work on the dead reckoning abilities of ants, and Gould's & Gould's observations of digger wasps' behavioral rigidity.[8] However, some insect behaviour goes beyond innately fixed action patterns, and implies that even the simplest forms of life have a "simple belief/desire psychology" in which perceptual systems work in tandem with belief- and desire-generating systems and some form of working memory to coordinate action schemata and direct motor control. Carruthers focuses on a number of studies on the communicative nature of honeybee 'dancing' which suggest that

> although basic bee motivations are, no doubt, innately fixed, the goals they adopt on particular occasions (e.g. whether or not to move from one foraging patch or another, whether to finish foraging and return to the hive, and whether or not to dance on reaching it) would appear to be influenced by a number of factors (Seeley, 1995) [...] Bees don't just accept and act on any information that they are offered [...] they *evaluate* it (Carruthers, 2006: 72).

Bees have distinct information states and goal states "which interact with one another in ways that are sensitive to their contents in determining behavior" (Carruthers, 2006: 78).

[8] I will not attempt to reconstruct these findings here, as they are beyond the scope of this chapter. Suffice it to say that Carruthers puts forward a fairly exhaustive survey of recent experimental data regarding invertebrate psychology which he returns to throughout his argument (Gallistel, Gould & Gould, Seeley, Menzel), and which I will only mention in passing, without directly referencing. Any conclusions drawn from the research mentioned in this paper are those of Carruthers, unless stated otherwise.

Carruthers further notes research by Tarsitano and Jackson (1994,1997,1998) on jumping spiders which displays elements of advance planning in hunting techniques – the spiders show clear indications of mentally mapping the most effective route to their prey, and then follow the selected route by memory, making adjustments on the go for the different spatial perspectives that would intervene during the actual journey (Carruthers, 2006: 78). These findings are all suggestive of a belief/desire architecture (in spiders and bees at least) as diagrammed below: (Carruthers, 2006: 66)



On this view, innate fixed-action patterns can be directly triggered by perceptual systems and routed via action-planning systems to motor-control. However, there are alternate pathways involving belief and desire generating systems which can also factor into action-planning and thereby allow such creatures to exhibit certain *thought-like* behavior worthy of the designation of 'concept-users'.

If even simple invertebrate animals can be proven to possess some form of simple belief/desire psychology, then "it would surely be quite extraordinary if belief / desire / planning architectures weren't extremely common in the animal world" and hence would be a major underpinning of the cognitive inheritance that mammalian life should expect (Carruthers, 2006: 66). Carruthers takes this foundation and builds slowly upon it with an analysis of mammalian psychology that paints a clear picture of how it could have been deepened and expanded to explain even the complex cognitive behavior of primates (and by extension, humans). It is not difficult to imagine how simple, generic belief-generating and desire-generating systems could multiply and deepen via selective pressure. By the time we get to mammalian life, Carruthers argues, experimental observations suggest we can likely posit belief modules for at least foraging, causation, geometry, object properties, and number (Carruthers, 2005a). Furthermore, "each modular system presumably has some sort of domain-specific memory function attached" (Carruthers, 2005a: 74). This "deepening" of existing modules via evolution is a key point for Carruthers, and it seems quite intuitively plausible, as described below:

> Thus some sort of social relationships module gradually developed into the beginnings of a mind-reading module; the foraging module became increasingly sophisticated, developing into a system of naïve biology; the causal reasoning system developed into a form of naïve physics; the object property system expanded greatly to allow for many more object categorizations; and so on (Carruthers, 2005: 75).

With an assumed sub-structure of belief/desire psychology which stretches deep into evolutionary history, it would arguably take only a little co-opting, expanding, deepening, and interconnecting of the pre-existing systems to end up with a massively modular belief / desire / planning cognitive architecture that appears to *at least* be present up to and including our great ape lineage, given experimental evidence.

At this point in his argument, Carruthers spends some time discussing the "dual vision system hypothesis", as it contains a number of assertions which will figure prominently in his full account of how the MM mind could function. He uses the experimental evidence gathered by Milner & Goodale (1995), Jacob and Jeannerod (2003), and Glover (2004), as well as Weiskrantz's (1986, 1997) research into the phenomenon of 'blindsight,' (Carruthers, 2006: 87) to suggest a model of a dual primate visual system in which visual perceptions are processed in both a ventral temporal lobe system and a dorsal parietal lobe system. The outputs of the former are "globally broadcast", in the sense that they are made available as *inputs* for other modular systems (such as belief-generating modules) concerned with object recognition, memory formation, and action planning; whereas outputs from the parietal lobe visual system are routed directly to action-schemata and put through to motor-control. The parietal lobe system operates beneath the level of "consciousness", and is responsible for "on-line visual guidance of movement" (Carruthers, 2006: 84). The subdoxastic functioning of the parietal guidance system is an important point, as on-line visual guidance would require extremely *quick, reflexive* operation, while the temporal system, operating in parallel, can afford to move more slowly (and *thoughtfully*, as it were), since "thinking while acting [can have] a detrimental effect on skilled performance" (Carruthers, 2006: 87).[9] Additionally, these visual systems are not entirely "feed-forward" in their operation, but "in fact, both systems contain very substantial back-projecting neural pathways."

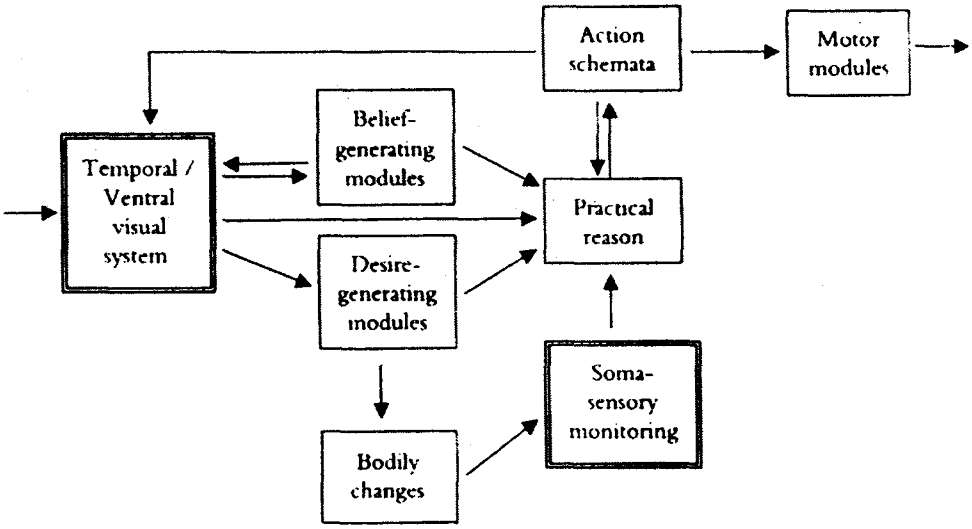
> The functions of the back-projecting pathways in the ventral system [...] are used to direct *attention* of various sorts towards aspects of incoming information [...] to 'query' the perceptual input, helping to resolve the interpretation of degraded or ambiguous input [...] The patterns of neural activity created earlier on in the ventral stream by the activation of these back-projecting pathways are then processed in the normal feed-forward manner [...] giving rise to a conscious experience *as of perceiving* (Carruthers, 2006: 92-93).

The idea here is that *cycles*, or feedback loops can be created in which representations are re-fed through the ventral visual system, in ways that can help with object recognition, for example, but which could also go a long way to explaining how objects could be *mentally* perceived and transformed, such as through mental rotation. It also offers a glimpse at

[9] Also note the well known highway-driving phenomenon of arriving at a destination safely with *no recollection* of having consciously controlled the vehicle could be easily explained by the work of a dual visual system, with the parietal system doing the lion's share of the work subdoxastically.

where Carruthers is going with this argument: that the existence of back-projecting pathways *in other systems* and the *global broadcasting* of outputs from the ventral visual system could offer a fruitful explanation of how motor-control systems could be activated *without* a direct perceptual inputs to trigger them, allowing for a sort of *mental rehearsal of action schemata.*[10]

Carruthers updates his former belief/desire/planning diagram to reflect the existence of back-projecting pathways in the connections of the ventral visual system to belief-generating modules and fixed action-schemata: (Carruthers, 2006: 141).


Action schemata
Motor modules
Belief-generating modules
Temporal / Ventral visual system
Practical reason
Desire-generating modules
Soma-sensory monitoring
Bodily changes

[10] This notion of global broadcasting of modular outputs so that they can then be taken up (or re-taken again) as inputs by other (or the same) modules forms a key part of Carruthers' model, and will be elaborated on in much more depth below. For now we will note two things which will be taken up separately as the discussion moves forward: one is the comparison to Baars' concept of the "global workspace" model of cognition in which memory works as a kind of "central information exchange" or "blackboard" via which a distributed collection of dedicated processors can communicate with one another, and without any central executive (Baars, 1988: 87). Parallels to Baars' work will be drawn throughout this exegesis of Carruthers and later on when I move into a more direct critique of Carruthers' model, as Baars has some insights into contextualization that may help Carruthers out of some difficulties I will raise for him in section 2.3 and attempt to answer in chapter three.

The second note about Carruthers' "global broadcasting" is that as a model it offers itself up to empirical verification, as if it is ubiquitous as Carruthers makes it out to be in the argument that follows, then certainly it should be something that experimental evidence should be able to pinpoint, at least in theory. Whether there is any evidence for it in the experimental literature will be a focus of chapter three of this thesis.

With an architecture wired in this way, one can see how Carruthers will make the case that potential action schemata can be mentally rehearsed and then fed back through the system to trigger *new* beliefs or desires, which can then be monitored (via some sort of *somasensory monitoring system)*[11] to provide further cognitive or motivational results. This fragmented, interconnected cyclical system should be taken *as a whole* as a sort of de facto practical reasoning system:

> [T]here *is* good reason to think of the set of action-controlling modules as collectively constituting an overarching practical reasoning module. For there is now more than just competition amongst the components. Rather, there are general mechanisms for adjusting motivation levels up and down in the light of somasensory information resulting from mentally rehearsed action schemata (Carruthers, 2006: 146).

The argument so far seems quite solid insofar as it posits a massively modular mammalian cognitive architecture which is capable of the primitive forms of 'creativity', mental mapping, and mental rehearsal of action that have all been observed in experimentation with animals.[12]

Thus far, a Fodorian objector may well have no difficulties with Carruthers' account, as no one is making a serious argument that primates are capable of the kind of holistic, abductive reasoning that humans are. So, this sets us up for the central question of this section: what changes take place between the upper primates and *Homo sapiens* which can account for abduction while still maintaining a massively modular structure, and *without* the addition of some sort of domain-general, content-neutral systems? Fodor, of course, thinks

[11] The details of such a system, and whether it is plausible to posit one are issues that will be explored in the critique of Carruthers' account later in this chapter. For now, suffice it say that the assumption of a somasensory monitoring system may spell some trouble for Carruthers, as it may require some sort of holistic central executive that is capable of doing the "monitoring" which implies some sort of global outlook and a wide grasp of the background somasensory state, not to mention an ability to make judgments or at least measurements. To posit such a monitoring system may be sneaking the ghost in the back door of the machine.

[12] See the literature on the chimpanzee Belle (Menzel, 1974), which is discussed at length in sections *2.3* and *2.8* of Carruthers (2006) for a wealth of examples of primate mental rehearsal of action. A full discussion of this data is beyond scope of this chapter, but it quite nicely illustrates behavior that fits perfectly with the account Carruthers gives of primate cognitive architecture.

there is something quite *special* and *unexplained* about human cognition which is not found in primate cognition, and won't be accounted for by simply adding more modules. He even argues the standard line about how close we are genotypically to our primate cousins, noting that "[o]ur brains are, at least by any gross measure, very similar to those of apes; but our minds are [...] very different" (Fodor, 2000: 88). Carruthers disagrees on both points: first, he points out that the argument "that we share 98.5% of our genes with chimpanzees [...] vastly underestimates the genetic differences."

> One reason is that even when genes are indistinguishable, they can be spliced differently during the process of transcription [...] When insertions and deletions are also included, the differences between humans and chimpanzees are much more significant, yielding a figure closer to 87% in common [...] And when one looks specifically at sequences of DNA known to be involved in gene regulation ['junk' DNA], what emerges is that the differences [...] are of the order of 15% (Carruthers, 2006: 153).

Carruthers would also disagree that our *minds* are actually so different, as on his view, human cognitive architecture only needs the addition of one altogether new module (language) and a deepening and expansion of co-opted existing primate modular systems in order to account for 'distinctively human thinking', as shall be explained in the remainder of this section.[13]

It is important to note that Carruthers takes great pains to disclaim that he is *not* arguing that "one new adaptation" (language) is sufficient to explain human cognition – in fact, he enumerates 22 aspects of 'distinctly' human cognition which must be accounted for, abduction included (Carruthers, 2006: 154-157) -- but the language module is certainly the centerpiece of the adaptive jump from primate to human thinking. His main argument can be summed up as suggesting that

[13] For a succinct explanation of the specific role the coming online of the language faculty serves in Carruthers' model, see *Consciousness* (2005), in which he sets out a seven point "argument for the claim that conscious propositional thinking is conducted by means of natural language sentences" (Carruthers. 2005b: 117-118). The discussion of these ideas that follows in this thesis will be focused purely on the ways in which natural language can support a modular account of holistic reasoning, and not the further argument that it is constituitive of conscious thought in general, though the argument is of tangential interest.

> distinctly human 'general intelligence' *[g]* involves mental rehearsal and the global broadcast of perceptual and quasi-perceptual (imagistic) information, as well as the representation and development of imaginary scenarios [...] My own suggestion is that *g* is a special sort of *interaction* effect of existing modular systems, together with some evolutionarily novel dispositions and tendencies. One aspect is that a pre-existing capacity for action-rehearsal utilizes the resources of the human language-production module to broadcast representations of sentences globally to the full range of central/conceptual systems, initiating cycles of inner speech (Carruthers, 2006: 166).

In terms of the "evolutionarily novel tendencies and dispositions" mentioned above, some of these take the form of modular systems distinct to humans, the existence of which is buttressed by copious amounts of experimental data. The human capacities of folk physics, folk biology, and folk psychology (mind-reading) can all be explained as deepenings of the pre-existing primate modules regarding object properties, causation, intentions of others, etc (Carruthers, 2006: 167-174). The bulk of the 'distinctly human' cognitive characteristics emerge from the "interaction effects" that the addition of the language module makes possible.[14]

Before we get to an explanation of how that works, however, there is an important issue which needs to be addressed regarding whether *any* of these "evolutionarily novel tendencies" are still to be assumed as the products of a *modular* architecture. Could not the evolution of a single domain-specific general learning device account for the differences between chimps and humans? Gopnik and Meltzoff, for example, present such an argument with their account of cognitive evolution based on the analogy of the little scientist (1997).[15]

[14] cf. Baars (1988, pp.87-88); see footnote 19 below for more.

[15] Segal (1996) has perhaps the cleanest take-down of Gopnik and Meltzoff's account of general learning mechanisms, pointing to the abilities of patients with William's Syndrome who, despite often severe mental retardation, show surprising facility with language. Such individuals clearly prove that language acquisition cannot be a result of a general learning device, according to Segal.

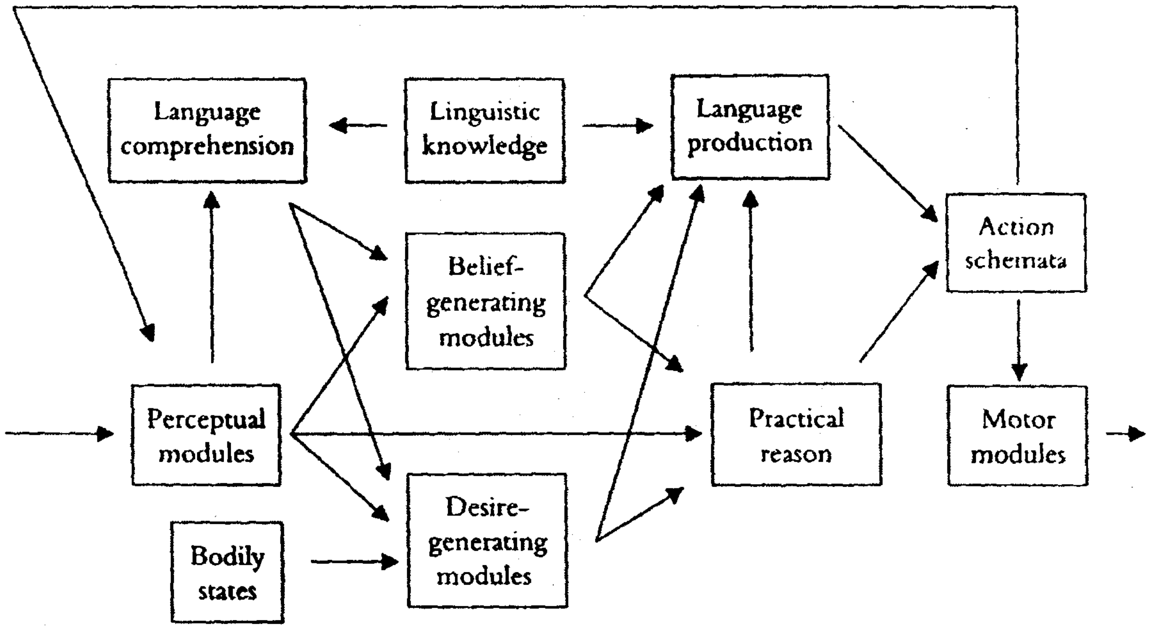
Even if it were settled that a general learning mechanism is unlikely to be the source of our "evolutionarily novel tendencies," there exists a further question regarding whether the 'modules' of folk biology, physics and psychology actually merit the name *module?* Are they 'encapsulated' and 'domain specific' in the sense that Fodor, for example, would define those terms? As Carruthers has defined the terms 'encapsulated' and 'module', everything that he is saying about these hew distinctly human 'modules' is consistent with calling them such. A critical discussion of whether or not Carruthers' terminological definitions are fair, whether they coincide with the definitions Fodor uses, and whether they are *fruitful* as definitions will all be set aside until the next chapter. In the meantime, at least in terms of a 'language module' it would seem that Carruthers is on solid ground with respect to the Fodorian definition of module, "for language forms one of the archetypical input and output modules defended at length by Fodor" (Carruthers 2003: 4). Indeed, to cite Fodor directly:

> When we look at real, honest-to-God *perceptual* processes, we find real, honest-to-God informational encapsulation. In parsing, for example, we find computational mechanism with access only to the acoustics of the input and the body of 'background information' that can be formulated in a certain kind of grammar. That is why [...] there are no context effects in parsing. AND IT IS ALSO WHY THERE IS NO FRAME PROBLEM IN PARSING (Fodor, 1987: 141).[16]

Carruthers' proposal of how human intelligence results largely from the coming on-line of language and mind-reading modules is quite complex, and carefully constructed. A full explanation obviously being beyond the scope of this section, we will have to suffice with a somewhat truncated breakdown. Taking for granted the previously discussed belief/desire/planning architecture, including back-projected wiring pathways and global broadcast of some outputs (such as those of the ventral visual system), Carruthers argues that the addition of a language-production system that can avail itself of the globally

[16] Jackendoff, for one, agrees that language perception is modular, though his views diverge from Fodor in interesting ways which will be elaborated in more detail in the next chapter.

broadcast outputs in the architecture allows the mind to engage in *cycles of inner speech*, as diagrammed below: (Carruthers, 2006: 233)



The language production system can produce sentences which *do not necessarily* have to be actually routed to motor control and spoken aloud. The global broadcasting of outputs from that system could feedback into belief and desire-generating systems, causing novel motivational states, which trigger new action schemata, which in turn trigger new sentences in language production, and so on. And because the outputs of the language system are (like those of the ventral visual system) *conscious*, they give us the inner phenomenon of *talking to ourselves* – the mental rehearsal of speech action schemata – and these account for our conscious "thoughts". Although there are thinkers who argue differently, it is generally assumed that we do not think in *natural language* of course, but in some sort of *mentalese*.[17] The important point here is that the language production system, if it can avail itself of the various mentalese outputs of other systems, can serve as a sort of universal translator, and in essence bring some of what would otherwise (in simpler life forms) be subdoxastic to a doxastic 'conscious' level.

Language gives us access to our thoughts, and the structure of grammar in the language system allows for the construction of *novel* sentences, the changing or substitution of components of existing sentences in the mind's speech action schemata database (which is constantly growing via memory), and the capacity to *conjoin* and *integrate* contents (Carruthers, 2006: 240) – opening up *variables* for cycles of inner speech, and the *creative* aspects that this variability and potential for novelty entail (Carruthers, 2006: 307). This is a point which will be returned to in the next section of this chapter, discussing how Carruthers' account can explain creativity and abductive inference. For a succinct summary of the argument, see Carruthers' "On Fodor's Problem" (2003):

> There is good reason to think that this [language] module would have been set up within the architecture of a modular mind in such a way as to take inputs from all of the various conceptual modules, so that their contents should be reportable in speech. And there is reason to think that the abstractness and re-combinatorial powers of natural language syntax would make it possible for the language faculty to combine together sentences encoding the outputs of different modules into a single natural language representation. If such sentences can then be displayed in auditory or motor imagination, they can adopt some of the causal roles distinctive of thought, then we shall have explained how thought can acquire some of its flexibility of content within a wholly modular cognitive architecture (Carruthers, 2003: 508).

[17] Or, more specifically, "multiple mentaleses" as Carruthers notes that mentalese cannot be thought of as some "lingua franca of the mind" (Carruthers, 2006: 51). The actual relationship between mentalese and natural language is the subject of much debate, and generally revolves around 4 specific theoretical axes, according to Steinberg et. al.: 1) that speech is essential for thought; 2) that language is essential for thought (but not necessarily speech); 3) that language determines or shapes our perception of nature; and 4) that language determines or shapes our world view (Steinberg et. al., 2001: 246).

This account may draw immediate fire, however, insofar as it may explain how the process of creative 'thought' may *proceed* in cycles of inner speech, but it doesn't necessarily explain how the process *starts:* novel sentences can *in principle* be generated via grammatical structure, and *once generated* can feedback into the cycle, but *how are they generated* in the first place? Carruthers has a number of suggestions for how this could work. First, there is the possibility that some mentally rehearsed sentences can be generated literally at random, modeled on the "'Protean' erratic behavior" of animals (even invertebrates), who almost all "have ways of activating sequences of [unfixed] action schemata that *aren't* determined by prior thought and planning" such as, for example, the randomized flight patterns of moths trying to escape a predator (Carruthers, 2006: 136). Second, certain rehearsals of speech actions can be triggered via the associative process of *implicit memory priming* -- both "perceptual", or modality-dependent priming and "conceptual", modality-independent priming (Carruthers, 2006: 126). Finally, these two elements come together in the distinctly human phenomenon of imaginary play. Many mammals 'play' in the sense of 'acting out' certain action schemata learned via imitation and direct instruction, and this play often serves to fine tune and reinforce those action schemata for when they will be utilized 'for real' in adult life (hunting, fighting, etc.) In human children, play is often divorced from specific 'skills' and generally involves communication with imaginary objects and interlocutors, yet, importantly, *outside of any actual communicative context* – it is 'practice'. Carruthers suggests that the *outward*, out-loud imaginary play of children, is in a sense a mental rehearsal of action schemata (mostly speech actions), outside of any *actual* communicative context, and that as the child matures and the mind-reading system comes on-line, that imaginary play moves *inward*, and takes place in cycles of inner speech (Carruthers, 2006: 304-309).

> [C]onsider the case of a young child pretending that a banana is a telephone. The overall similarity in shape between the banana and a telephone handset might be

> sufficient to activate the representation TELEPHONE, albeit weakly. If the child has an initial disposition to generate an appropriate sentence from such activations, then she might construct and entertain the sentence. 'The banana is a telephone'. This is then comprehended and processed, accessing the knowledge that telephones can be used to call people, and that grandma is someone who has been called in the past. If the child *likes* talking to grandma, then this may be sufficient to initiate an episode of pretend play. By representing herself *as* making a phone call to grandma (using the banana), the child can gain some of the motivational rewards of a real conversation. The whole sequence is thus reinforced [...] From such simple beginnings, one can imagine that children gradually build up a set of heuristics for generating fruitful suppositions [...] leading eventually to the capacity for creative thinking and problem solving which is distinctive of human adults (Carruthers, 2003: 511).

The key point is that all the hallmarks of creative thought are the same as those of explicit imaginary play in childhood: associative memory priming plus some random variability allowed by grammatical syntax allow for novel, creative sentences to be entertained, novel speech action schemata to be mentally rehearsed, which are broadcast globally and can have effects on other belief- and desire-generating systems. In short:

> action schemata for items of speech can be assembled in the absence of any prior thought-content for them to encode, but rather for purposes of supposition. We can *try out* saying things, either out loud, or in inner speech, using various heuristics for this generation (Carruthers, 2006: 311).

Note that this is a much fuller elaboration of the 'supposition generator' that Carruthers' earlier accounts posited (2003, 2005a), but left arguably underexplained. The *supposer* is an interaction effect, rather than a module in its own right, which is likely more plausible than the notion that a supposition generator could evolve as a distinct mechanism.

The mind-reading system also plays an important role here, insofar as when combined with the language production faculty, the mind-reading capacity can be turned *inwards*, allowing the subject to 'read his/her *own* mind' as it were. Carruthers notes extensive research into confabulation and self-ascribed theorizing which has found "powerful evidence that we do actually attribute beliefs and goals to ourselves as a result of swift and unconscious *self-interpretation*" and that "many of our beliefs about the thought processes that cause our own behavior are actually *confabulated*" [...]

> Far from having direct access to our own reasoning and decision-making processes, what we actually do in many cases is interpret our own behavior, *ascribing mental states to ourselves in much the same sort of way that we might ascribe them to another person.*[18] And where the true causes of behavior are obscure to common sense psychology, such self-attributions will frequently be false (Carruthers, 2006: 179, emphasis mine).

The well known observations of Gazzaniga regarding his split brain patients bear this out.

> Human brain architecture is organized in terms of functional modules capable of working both cooperatively and independently. These modules can carry out their functions in parallel and outside of the realm of conscious experience [...] Monitoring all of this is a left-brain-based system called the interpreter. The interpreter considers all the outputs of the functional modules as soon as they are made and immediately constructs a hypothesis as to why particular actions occurred. In fact the interpreter need not be privy to why a particular module responded. Nonetheless, it will take the behavior at face value and fit the event into the large ongoing mental schema (belief system) that it has already constructed (Gazzaniga, 1988: 219).

Additionally, one of the most documented psychological phenomena is our uncanny ability to lie convincingly to ourselves regarding motivations and intentions to *a posteriori* force accordance with subsequent behaviour, such as the seminal work of Festinger whose "cognitive dissonance" experiments (1957) proved that we are all like Aesop's fabled fox who never wanted those sour grapes anyway. Festinger's "counterattitudinal" experiments showed that subjects would shift their attitude about a particular unpleasant task in a more positive direction if they were paid less for it, as it was psychologically problematic (dissonant) to believe that one had performed an unpleasant chore for little return, so the belief about the pleasantness had to be revisited and revised (Festiger, 1957; cf. Carlsmith,

[18] This notion leads to many interesting possibilities regarding the role that a theory of mind module might play in the framing of a *conscious sense of selfhood*, which I will leave aside for now, but return to in the final chapter. One of the by-products of a cognitive model such as Carruthers', if it is successful, is that it may accidentally hit upon an answer to the notorious "hard problem" of consciousness (Chalmers, 1995) in the sense that we frame *ourselves* by treating our selves as if we were minded, in the same fashion that we attribute mind to others thanks to the folk psychology faculty. It would be a strange sort of Cartesian inversion, where one might say "S/he thinks, therefore I am." The self as an analogue of the other.

1959).[19] It is clear to see that Carruthers is likely justified in positing a faculty of mind turned inward as the mechanism via which self-ascription of motivation and self-interpretation of behaviour may be made.

It is not Carruthers' contention that *all* self—interpretative processes are "epi-phenomenal", but rather that "much *less* of our behavior may actually be caused conscious thought-processes than we are intuitively inclined to believe" (Carruthers, 2006: 180).[20] Carruthers moves from this assertion to a postulation that a full explanation of human performance will necessitate "two *sorts* of reasoning systems" --

> a set of swift unconscious reasoning modules, on the one hand, and laborious conscious reasoning in accordance with believed normative standards, on the other. And the latter, utilizing globally broadcast sentences in 'inner speech', as well as other forms of imagery, requires only a mind-reading system that has access to perceptual input (Carruthers, 2006: 185).

This 'dual reasoning system' (closely allied with the work of Frankish)[21] mirrors the dual visual system in that it explains how a great deal of the subdoxastic work can go on, while

[19] Interestingly, Festinger's account of cognitive dissonance relies also on a mental ability to judge individual thoughts are relevant to other thoughts – only relevant thoughts can be dissonant with one another. Festinger does not discuss framing issues, or any mechanism via which this "relevance" could be judged, but we will return to this point in the third chapter, as dissonance theory may hold some insights into context framing and relevance determinations with regard to Sperber's views.

[20] This line of argument also ultimately leads Carruthers to align himself with Wegner to a certain degree, in arguing that the idea of 'conscious will' is somewhat of an illusion. A full discussion of this is beyond the scope of this chapter, but I will note the conclusion here, as it is of tangential interest to the argument about globality and whether our 'holistic' reasoning is truly global, or just an illusion of globality:

> "Wegner (2002) is correct: conscious will is an illusion. Given that the mind-reading system has no direct access to events that take place in the practical reasoning system, but only to their globally broadcast effects, then our access to those events is always interpretative. Hence those events within practical reason don't qualify as conscious ones [...] Only if I want to do what I have decided and *believe* that by saying to myself, 'I'll do Q', I *have* decided, does the action get settled upon" (Carruthers, 2006: 412).

We will return to the topic of consciousness and how it may relate to this modular account of cognitive architecture in the third chapter.

[21] See Frankish's *Mind and Supermind* (2005) for a full elaboration of his account of *system 1* (subpersonal/non-conscious) and *system 2* (personal/conscious) reasoning. A fairly clear summary of this account is found in Carruthers, 2006, pp.374-382.

still leaving a parallel architecture for the slow, linear, "laborious" work of 'conscious' reasoning which we call distinctly human. This distinction of these two systems will play an important role in explaining how Carruthers' account can answer the objections of Fodor, as we shall see in the following section.

2.2 Carruthers vs. Fodor's 'input problem'

Fodor was quite clear about what he thought a massively modular architecture would be unable to account for: abductive, holistic reasoning. Holistic reasoning would necessarily depend on an ability to query the entire backdrop of prior epistemic commitments, which would entail exhaustive searches which would be computationally intractable. And any attempt to limit those searches and impose frames or 'relevance constraints' via either local heuristics, or dispersed, encapsulated, domain-specific modular systems is doomed by the *a priori* 'input problem' of how representations get 'tagged' and routed to the appropriate local systems. Remember that, according to Fodor, any local domain-specific system *presupposes* a more global, less domain-specific system via which representations are assigned as inputs to various modules. So, the question which concerns us here is whether Carruthers' model of MM cognitive architecture can get around these objections and do what Fodor claims it couldn't possibly: account for abduction.

On Carruthers' view, there is "one overarching decision-making / practical-reasoning system (albeit made up of several sub-modules), and this will be the point at which 'everything comes together'" (Carruthers, 2006: 225). Everything 'comes together' in practical-reasoning insofar as the outputs of various belief- and desire-generating systems are all made available to it via global broadcasting, and cycles of inner speech. The practical-reasoning system is *not* a single module, but an "interaction effect", so in this sense it doesn't suffer from an "input problem" in the Fodorian sense – it is essentially domain general in terms of inputs. A practical reasoning system viewed in this way does, however, propose immediate tractability concerns, which Carruthers notes, since "the decision-making system is the point of maximum convergence of information [...] there will therefore be maximum demands on the computational resources of the practical-reasoning system" (Carruthers, 2006: 225). Fodor would certainly expect this combinatorial explosion, especially since the supposition is that Carruthers' overarching practical reasoning system receives *all* outputs from other systems as inputs (including, presumably, their attendant *memory* systems), and would have exhaustion issues.

> Fodor might respond that there is no way of knowing *in advance* which of one's beliefs might fail to cohere with a hypothesis. In which case, he might say, we have no option but to consider them all if we are ever warranted to make an inference to the best explanation [...] But something less demanding and more heuristic-like will have to suffice (Carruthers, 2006: 362).

Carruthers' answer to this is the deployment of heuristics – "'quick and dirty' shortcuts, in order to ease computational load, and to render [the] task more frugal" (Carruthers, 2006: 362). Specifically, Carruthers suggests we use three distinct types of heuristic rules: 1) those concerning *how long* a search among alternatives should be entertained before making a decision; 2) those concerning *what sorts of information* are relevant in the given context; and 3) those concerning *which action-schemata* in one's database to activate for mental rehearsal (Carruthers, 2006: 228-229). All of these heuristics are an attempt to frame contexts, limit searching, and, in a sense, *impose* relevance.[22] Carruthers sides with Sperber and Wilson (1996, 2002; also Newell & Simon, 1990), arguing that we generally "adopt a *satisficing* policy,

[22] Note also that the three heuristics Carruthers cites here line up perfectly with the three aspects of the frame problem that Fodor highlights, *where to search, how long to search, and what the nature of the initial inquiry is.* As discussed in chapter one, a doctrinaire computer scientist would say these are three very different problems and should not be conflated as *the* frame problem in the way Fodor seems to (as Hayes objects).

governed by heuristics [...] If they achieve results that are relevant enough, they stop" (Carruthers, 2006: 370).[23]

Fodor would have none of this talk of heuristics somehow paving the way to abductive inferential reasoning of course. Certainly, heuristics can limit searches and provide tractability, but a variation of the 'input problem' comes to bear of the *selection* of heuristics. Fodor argues:

> It is circular if the inferences that are required to figure our *which* local heuristic to employ are themselves often abductive. Which there is every reason to think they often are. If it's hard to model the impact of global considerations in *solving* a problem, it's generally equally hard to model the impact of global considerations on *deciding how* to solve a problem [...] since deciding how to solve a problem is, of course, itself a species of problem solving (Fodor, 2000: 42).

For Fodor, this leads to a vicious regress, as one would have to appeal to heuristics for the selection of heuristics ***ad infinitum***, and the specter of non-modular abduction is still there at the end of the regress.

There are two very effective answers to this objection which can be drawn out of Carruthers' account. The first, and simplest, answer is that there is no reason that the selection of heuristics cannot be, to a certain extent, *arbitrary*. As long as the *satisficing* policy is assumed, then there doesn't seem to be any reason why the decision about *which* heuristic to use need be 'rational' in some ideal sense that implies total global consistency with all background beliefs. A heuristic like "take the last"[24] is completely arbitrary in it's operation,

[23] The notion of 'satisficing' is laid out clearly by Gigerenzer & Todd who define it as "the shortcut of setting of an adjustable aspiration level and ending the search for alternatives as soon as one is encountered that exceeds the aspiration level" (Gigerenzer & Todd, 1999: 13).

[24] Gigerenzer highlights a number of basic heuristics in our "adaptive toolbox" including "take the last" – a selection heuristic which directs an agent to (arbitrarily) select the same action as the time before, if that strategy worked the last time, it may very well work again (Gigerenzer & Todd, 1999). An example of this heuristic in action is in the choice of what to eat at a particular restaurant. A simple and effective way to limit the choice (assuming the menu is gigantic) is to simply order the same thing as last time. It is an arbitrary choice, but may nevertheless be satisfying, and probabilistically more *likely* to be satisfying than an another, untested choice.

and may very often be employed in an automatic, non-rational-seeming fashion – yet it may still *satisfy* the subject employing it, insofar as it prompts action schemata that satisfy the underlying motivations as put forward in the subject's belief/desire psychology. Samuels makes a good point along these lines regarding the (seeming) globality of reasoning which is relevant to this distinction, when he argues that

> it is important to keep firmly in mind the general distinction between normative and descriptive-psychological claims about reasoning: claims about how we *ought* to reason and claims about how we *actually* reason (Samuels, 2005: 118).

Just because ideally 'rational' reasoning about, for example, which heuristic to use in a given context *should* be global, doesn't mean that we *actually* satisfy the exhaustive demands of that globality in the actual process. Indeed, it seems as if we do *not*, especially when it comes to selecting heuristics, which seem to be selected from a fairly *short* list, which doesn't require a lot of background searching for relevance at all. Carruthers pushes for a more "naturalized rationality" (cf. Stein, 1996), that "what is rational for us depends upon our powers and limitations" (Carruthers, 2006: 361). Pinker argues this also, in his direct response to Fodor, when he contends that "people use heuristics" which result in "statistically useful but frequently fallible information" (Pinker, 2005: 15).

Gigerenzer and Todd have a great deal of helpful things to say on this type of naturalized rationality, or "ecological rationality" as they prefer: "rationality that is defined by its fit with reality [...] We propose replacing the image of an omniscient mind computing intricate probabilities and utilities with that of a *bounded* mind reaching into an adaptive toolbox filled with fast and frugal heuristics" (Gigerenzer & Todd, 1999: 5). This view of bounded rationality does seem to jibe with the way people *actually* think, since no human mind actually possesses a demonlike unbounded rationality, despite the fact that we may believe and *act* like we are unboundedly rational (Gigerenzer & Todd, 1999: 9). In their account, Todd and Gigerenzer do recognize the difficulty Fodor raises about *how* certain heuristics are selected, or *what* selects them, noting that "the fact that heuristics are designed for particular tasks rather than being general-purpose strategies solves part of the selection problem by reducing the choice set[25] [...] but we have not addressed how individual heuristics are selected from the adaptive toolbox for application to specific problems" (Todd & Gigerenzer, 1999: 364-365). However, no matter what, the list of available heuristics, if they are part of an adaptive toolbox, will not be nearly as unbounded as Fodor's objections suggest. The ostensible globality of the heuristic selection process assumes a requirement of unimpeachable *correctness* of the choice. Yet, for the naturalized rationalist, to a certain extent, if it works, it works. Heuristics that did not generally work would never have been selected for and therefore would not even be in the toolbox (no matter how fast or frugal they may be). As Goodie et al note:

> The frame problem illustrates one reason why bounded rationality must be applied in a domain specific way. While the techniques of heuristic search applied widely in AI are intended to reduce the need for demonlike computational capabilities, they often fail in this mission because the heuristics involved, while fast and frugal, are not fit in the environment on which they are unleashed (Goodie et al, 1999: 333).

Fodor would certainly interject at this point that the *truth* of the belief that gets fixed via reasoning should play some role, and that 'if it works, it works' doesn't respect that fact. As he writes:

> data isn't useful unless it's true; and only instructive processes can yield true data reliably and on a large scale [...] and since the world is prior to the mind, there is no way that the required epistemic correspondence between the mind and the world can be achieved unless the world can shape what the mind believes (Fodor, 2000: 96).

However, contra Fodor, it is quite well established that we consistently engage false beliefs -- many of which are adaptively advantageous, as Cosmides and Tooby very capably argue

[25] "External factors, such as time pressure and success may help to select heuristics. There are also certain situations in which it is adaptive to alternate between multiple strategies, either randomly, yielding unpredictable *protean* behavior [...] or systematically" (Gigerenzer & Todd, 1999: 32).

(Tooby, Cosmides, Barrett, 2005; also Pinker, 1997) which only seems to support the position on heuristic usage being laid out here.

The second, and perhaps more interesting answer for how heuristics are selected falls out of Carruthers' discussion of Barrett's (2005) 'enzyme account' of modularity, in which

> the idea is to model the operations of modules on the way in which enzymes build proteins within cells [...] Each has a characteristic shape, and floats around waiting to meet a protein that matches that shape [...] Translated into cognitive terms, the idea is that there might be a whole host of specialist processing devices, all focused on a common 'bulletin board' of representations. Whenever a device comes across a representation that 'fits' its input condition it gets turned on, and then it performs some set of transformations [...] before placing the results back on the bulletin board for other devices (Carruthers, 2006: 219).

Remember that in Carruthers' MM model, global broadcasting and cycles of mental rehearsals of action schemata (mainly 'inner' speech actions in humans) would constitute a sort of constantly evolving, interactive 'bulletin board'. And the enzyme account describes a process where certain devices (e.g. a heuristic selection device, perhaps) are constantly hunting for representations which 'fit' their inputs. In this way, the problem of *how* relevant heuristics are chosen is actually *inverted*: the heuristics are not 'chosen' by some decision-making system, but rather *choose themselves* by finding contexts they recognize as relevant to themselves, given their specific domain.

This also goes a long way to answering Fodor's *a priori* input problem (that the system via which representations get tagged as inputs *must* be domain-general, content-neutral in some respect). With the enzyme model in mind, we see that there is no homunculus which surveys the epistemic background and routes representations to appropriate target modules. Rather, the *modules themselves*, which are severely domain-limited, know very little *except for what they can take as input*, and they latch onto those inputs which trigger them. There is a slight danger here in personifying this process metaphorically, talking about modules "seeking out", "knowing", and "finding" appropriate inputs, which still seems to smack of a certain *decision* process (dragging us right back to a Fodorian objection). However, the process is *dumb,* not deliberative.[26]

Another analogy for how systems could 'find' the appropriate inputs, as opposed to the inputs being globally sorted, involves a sort of 'sifting' function, (which would be enabled by the cyclical broadcasting of outputs from various systems). Think of a coin sorting device in a vending machine: different sizes and weights end up triggering different openings so that the appropriate coins, after some jostling, end up in the appropriate cache, and the machine can 'count' the change and calculate the money fed into it. There is no *intelligence* involved in routing the coins to the appropriate inputs, excepting the intelligence of the *design.* But assuming the 'designer' is natural selection, then it seems entirely plausible to posit systems where inputs find their way to the appropriate destination *without* any globalized selection or deliberative process. There is the *appearance* of executive control here only.

Look back to the example of the 'cheater detection module' (CDM) which Fodor claimed had no way to discern the "very subtle clues" in the environment which would require its services. On Carruthers' account, it seems quite plausible that heuristics for finding *certain* very subtle clues could operate in order to tag some social contexts as those in which cheating is likely to occur. These would operate alongside associative memory

[26] Again, Baars comes in handy here with a useful metaphor:

> There is one especially apt analogy: a large committee of experts, enough to fill an auditorium. Suppose this assembly were called upon to solve a series of problems that could not be handled by any one expert alone. Various experts could agree or disagree on different parts of the problem, but there would be a problem of communication: each expert can best understand and express what he or she means to say by using a technical jargon that may not be fully understood by all the other experts. One helpful step in solving this communication problem is to make public a global message on a large blackboard in the front of the auditorium, so that in principle anyone can read the message and react. In fact, it would only be read by experts who could understand it, or parts of it [...] One effect of a global message may be to elicit cooperation from experts who would not otherwise know about it. Coalitions of experts can be established through the use of the blackboard (Baars, 1988: 87-88).

priming, as well as cycles of mental rehearsal of action which come to *statistically* determine the likelihood that a cheating situation was imminent. Fodor seems to think of the detection as surveying *all* the infinite detail of the scenario and drawing some inferences, whereas the enzyme account would have it that we are hardwired by natural selection to *focus in on* specific "very subtle clues." They only seem "very subtle" relative to the entire background; but to a device that *can only see* such things, it doesn't seem so difficult to "discern" them.

This concept of modules *seeking* appropriate inputs also explains how it is that some modules can backfire, insofar as they take *inappropriate* inputs sometimes, which are similar to the appropriate input in such a way as to confuse the module. Sometimes this will result in no processing, as the module can do nothing with it. Other times this will result in outputs that are incorrect (such as with optical illusions), since the device treats the input in a certain mandatory encapsulated way despite it not being the kind of input the device was designed to take.[27] This in turn can also add an element to any explanation of creativity or novelty of reasoning – since a mistaken input *may* end up providing a novel and *useful* output, which can then be broadcast and taken up by other aspects of the system, or harden a new and useful connection in place, selecting an existing mechanism for an entirely novel function – a cognitive spandrel.

2.3 *Carruthers on abductive reasoning*

So, Carruthers' model can explain some *seemingly* holistic / global cognitive processes and it can account for novelty and creativity in thought. But what of abductive reasoning in particular, which was Fodor's main concern? Obviously abductive reasoning contains

[27] For example, the way an octagonally shaped block might fit through a square hole in a toddler's game, or the way some relatively worthless Greek coin may fool a Montreal parking meter into thinking it has received a loonie.

elements of the above described processes, and in that sense we have already shown how Carruthers' architecture could potentially account for it, but he nevertheless takes the time to lay out a specific explanation of modular abduction in chapter 6 of *AotM*. His main focus is on the role of *testimony* in communication, and how it plays a role in a later development of abductive reasoning skills.

The mind-reading module was developed to judge the intentions of others, and with the advent of language, it also begins to play an important role, in connection with the language module, in determining whether or not to *believe* the testimony of others (whether what they say is meant to deceive or not would be a highly *relevant* detail in a hostile world!) Assuming over time we became quite adept at deciding whether to accept or reject testimony from *others* (based on such heuristic principles, presumably, as coherence with previously held beliefs, simplicity, etc.), then it is a simple jump to posit that a mind-reading module *turned inwards* to parse *inner speech* could use the same principles of testimony evaluation in order to judge *thoughts*. On this account, our capacity for abductive reasoning is arguably a *spandrel* of "other selected-for aspects of cognition (a modular language faculty, together with principles of testimony-acceptance and discourse-interpretation)" (Carruthers, 2006: 371). Carruthers concludes that

> a sort of *virtual* faculty of inference to the best explanation can be constructed from principles involved in the assessment of linguistic testimony and the interpretation of speech, leading to a preference for internally generated sentences that are consistent, coherent, and fit the data, as well as being simple, fruitful, and unifying (Carruthers, 2006: 386, italics mine).

Referring to a *virtual* faculty of inference suggests that Carruthers' account does not so much explain abduction, as explain it *away*. There is no type of optimal abduction inference going on, though it *appears* holistic, it is actually quite bounded. In this way, Carruthers is rejecting Fodor's terms of the abduction problem: we do *not* actually do what Fodor says modularity can't account for – we just *appear* to -- and what we *actually* do is much less optimal than 'inference to the best explanation'.

Carruthers' account, of course, is somewhat speculative and relies on two key assumptions: namely, that there exists a theory of mind *module* that is encapsulated and domain-specific at least to the degree necessary to frame the computational task of ascribing mind or intention to others (or by extension, to oneself), and that this module is capable of interfacing with the cycles of inner speech action schemata that Carruthers argues are constituitive of human thought. The question of whether these assumptions are well-founded will be addressed at length in chapter three. Ultimately, it is my contention that Carruthers' account of MM in *The Architecture of the Mind* can account for cognitive framing in a manner that does largely rebut Fodor's objections. However, there are a few potential weak spots which a Fodorian objector could exploit to try and pry open some holes in Carruthers' architecture. The final chapter of this thesis will address some of these possible areas of contention in order to assess the seriousness of the threat they may pose, and offer some lines of argument which could ameliorate and resolve the outstanding issues in a way that complements and supports Carruthers' core formulation.

CHAPTER 3:

Critiquing Carruthers' Architecture

I have argued that Carruthers suggests a fairly convincing model of a massively modular cognitive architecture that can successfully frame cognitive contexts in order to ensure computational tractability, answering the difficulties posed by Fodor without resorting to some sort of domain-general, content-neutral central processing executive. However, there are a number of potential weak spots in Carruthers' account where a determined critic could dig in and cast doubt on the entire enterprise. The focus of this final chapter will be on elucidating these potential oversights of Carruthers and arguing that it is indeed possible to get around them in a way that is both consistent with Carruthers' central argument, and satisfactory to his potential critics. Specifically, I will explore three possible objections to Carruther's position: 1) that he has employed a definition of modularity that is problematic and perhaps not germane to the issue of whether modular architecture can account for computationally tractable holistic thought; 2) that his account suggests certain empirical predictions which may be unfounded; and finally, 3) that Carruthers' is overly cavalier and sloppy in his reliance on 'relevance' and 'context' in his account of framing, setting himself up for the objections of circularity and question-begging.

3.1 *What counts as a module?*

The first line of possible objection to Carruthers' account would take aim at the definitions which underpin his entire concept of 'modularity' from the very start. Carruthers works from a definition of 'modular encapsulation' that he terms "wide-scope" as opposed to the strict, Fodorian, "narrow-scope" sense in which color perception or face recognition is deemed modular (Carruthers, 2006: 8-10). Furthermore, "there are a range of meanings of 'module' available," including, in his formulation, a 'module' that can be construed as constituted *by other modules*, interacting with one another, insofar as "a module can have other modules as parts" (Carruthers, 2006: 390ff). This is specifically *not* the sense of 'modular' that Fodor is working with when he argues that the MM thesis is flawed. On the contrary, Fodor notes that this conception in which "anything that is or purports to be a *functionally individuated* cognitive mechanism – anything that would have its proprietary box in a psychologist's information flow diagram – thereby counts as a module" would make *everyone* a modularity theorist outside of behaviorists and Gibsonians (Fodor, 2000: 56). Sperber too talks of the danger "when allowing for a great variety of modules networked in complex ways, of trivializing the notion of modularity to the point of confusing modules with boxes used in diagrams representing the flow of information in cognitive processes" (Sperber, 2005:56)[28]. The question is whether Carruthers, with his model of an *overarching* practical reasoning system has committed this slip, and has developed an architecture that is arguably *not* modular in the Fodorian sense. This is more than simply a definitional cavil, as the reason for employing modularity as an answer to the frame problem in the first place is that modular encapsulation obviates the need for frame axioms, as "only what is in [a module's] database can be in the frame [...] it doesn't need to treat framing as a *computational* problem" (Fodor, 2000: 64). If Carruthers' account is *misapplying* the term 'modularity' to processes which do not have the same computational restrictions as Fodor imputes to modules, then there may be no computational benefit to Carruthers' account simply by virtue of being titularly 'modular'.

[28] This objection by Sperber may turn out to be somewhat ironic, given the discussion in section 3.3 regarding how *he* appears to define modular processing.

Carruthers seems to be aware of this, and takes pains to argue that his "wide-scope" notion of encapsulation still allows the systems he so describes to be "modular", insofar as they still retain the claim of *mandatory operation* and are, for the most part, domain-specific (Carruthers, 2006: 9). The contentious part of any definition of 'encapsulation' lies in how one defines the access that the encapsulated processor has to anything else in the mind. In Fodor's strict sense, an encapsulated module is a sort of black box, whose operations cannot draw on any outside information, nor can any outside system intervene or even 'see' the processing – only the output. In Carruthers' 'encapsulated' module, this obviously has to be loosened, since modules can be sub-modules of overarching modules, and since modules can take globally broadcast outputs from other systems as inputs, which is *in a sense* a sort of bringing outside information to bear. A strict Fodorian would hardly agree with this looser formulation of encapsulation, and in that regard, one could argue that the entire debate laid out in this paper regarding the integratability of the MM thesis and abductive reasoning becomes illusory, as the two sides (Carruthers and Fodor) could be seen to be simply talking past one another.

This specter of talking past one another due to the (mis)use of shared terms without shared definitions is a constant problem in cognitive science.[29] It is precisely the same sort of objection, for example, that is made by Clarke against Fodor's attempts to refute Cosmides and Tooby: that Fodor is using a strict definition of module which Cosmides and Tooby never claimed to be using themselves, and hence isn't really attacking their position at all, but only a straw version of it (Clarke, 2004: 30). Clarke suggests that "for Cosmides and

[29] Fodor admits guilt in this regard himself, having expropriated the term 'module' from Chomsky's earlier work, but meaning something subtly different than Chomsky by the term, just as he is accused of having done to the 'frame problem' by Hayes, as discussed earlier. Now Fodor would argue the massive modularists have expropriated *his* terms and mean something quite different. Terminological misappropriation seems to be a running theme in cognitive science disputes – or at least disputes involving Fodor.

Tooby, the mind need not be *only* modular (or only massively modular). The mind, on their view, can also be nonmodular in certain respects [...] Their position, one might say, is fairly liberal concerning what might be the case" (Clarke, 2004: 7). According to Clarke, Fodor is fighting against a reading of Cosmides and Tooby that assumes they think the *entire* mind must be modular and therefore domain-specific, but that he is mistaken in doing so. Cosmides and Tooby, on Clarke's reading, are not against the idea that there may be domain-general, content-neutral structures in the mind – in fact he notes that they even suggest one: some sort of Bayesian inference module (Cosmides and Tooby, 1996) – but rather, they are only committed to the stipulation that the mind could not be *entirely* made up of such devices, for it could never have evolved to be so. This loosening of the demand for domain-specificity would indeed rob Fodor of his criticism that Cosmides and Tooby's model fails to account for abductive inference, but Clarke notes that Cosmides and Tooby do not in fact offer an explanation of *how* such inference might work even within the looser definition of massive modularity he ascribes to them. He argues that it is "compatible" but that "the fact remains that Cosmides and Tooby just do not have much to say about abductive inference" in terms of any positive argumentation (Clarke, 2004: 36).[30]

[30] Clarke's own account of abductive inference is that it "supervenes on means-end reasoning" (Clarke, 2004: 37). Citing much of the same experimental evidence that Carruthers uses also, such as the spatial reorientation work of Gallistel, Clarke posits a picture in which,

> it isn't explicit, occurent, transparent-to-the-cognizer abduction that occurs. Rather, it is subdoxastic, evolutionary abduction that predominates by means of natural selection [...] the moral is that much goes on beneath the surface [...] I think that a great deal of cognitive activity that looks like explicit abductive inference is really nothing of the sort. It is means-end reasoning by local, computational processors that were forged in the crucible of evolution by natural selection. Often, such reasoning produces beliefs that exemplify global abductive properties (Clarke, 2004: 38-42).

This account lines up quite neatly with Carruthers' version of how abduction is "virtual" in the sense that it *appears* global, but is essentially constrained in its computation. The mechanisms via which the seeming abductive inference supervenes on local computational structures is explained in slightly different ways, but both Clarke and Carruthers agree that abductive inference can be done within the massively modular mind, ***without*** resorting to some sort of domain-general, content-neutral central processor in the vein that Fodor suggests.

So if it is possible that Fodor's attack on Cosmides and Tooby could be mostly an attack on straw, then it is possible that the inverse is true in the case of Carruthers: that one could pick apart his claim to modularity and argue that he isn't saying much that Fodor would dispute – that his *answer* to Fodor's problem doesn't respect the precise terms of the problem. I will turn now to the work of two cognitive scientists, Ray Jackendoff and Simon Baron-Cohen, whose accounts of modular processing will be shown to support Carruthers' terminology, and the amenability of 'wide-scope encapsulation' to Fodor's original, narrower definition.

Jackendoff, interestingly, finds a way to widen the Fodorian definition of modular encapsulation by first *narrowing* it. He attributes to Fodor "the hypothesis that input and output systems are organized into faculty-sized modules" such as language perception (Jackendoff, 2002: 219). However, Jackendoff argues that "the locus of modularity is not large scale faculties such as language perception, but on the smaller scale of individual integrative, interface, and inferential processors" (Jackendoff, 2002: 219). This is to say that a language perception 'module' cannot be viewed as a single, encapsulated processor in the way Fodor describes it (although it *appears* as one). Rather, Jackendoff argues that everything we know about psycholinguistics suggests that a number of smaller, narrower modules are involved in language perception, along with some interface modules which allow lower-level syntactic modules to feed information to each other in such a way that integrated semantic meaning can be constructed from acoustic sensory information. One example that Jackendoff employs to illustrate this point is the strange but well documented 'McGurk effect' in which people shown a video of a person saying the syllable *ga* while the sound *ba* is overdubbed will report hearing *da* with no knowledge of the discrepancy.[31] In this situation, the visual information gained from watching the mouth form the syllable *ba* is integrated into the processing of the acoustic information, and actually *overrides* the output of the acoustic processor. Here is a clear case in which an ostensibly encapsulated module like the phonological parser (which Fodor himself claims is unambiguosly encapsulated) has its output *amended* by visual perception.

> Within structure-constrained modularity, the McGurk effect can be attributed to an additional interface processor that uses visual input to contribute fragments of structure to phonological working memory. But this interface can't tell phonology about all aspects of phonological structure – only about those distinctive features that can be detected by visual inspection [...] Similarly, its input is not all of visual structure, but only those aspects that pertain to the external appearance of the vocal tract. So it implements an extremely limited partial homology between the visual input and homological structure (Jackendoff, 2002: 225).

There is an interesting point being made about modular encapsulation here. Fodor holds up language perception, or at least *parsing*, as one of a quite short list of modular processes, yet here we see a clear violation of the ostensible encapsulation of the auditory parsing system. However, this is not to say that the parsing system is not in fact encapsulated – it merely proves the existence of some sort of bi-domain-specific, integrative, interface process which allows encapsulated modules to 'share' information. The parser does not really have *access* to visual perception in any true sense, and visual perception similarly cannot be said to have access to the parser, yet the faculty-level module of language perception obviously has at least some limited access to both. Jackendoff concludes:

> There is no extrinsic border around modules. Rather modules are *implicitly* differentiated, by what formats of cognitive structure they access and derive [...] Each module is strictly domain-specific in Fodor's sense: integrative and inferential processors deal with only one level of structure each; interface processors deal with two (we might therefore want to call them 'bi-domain specific'). Similarly, each

[31] Until they close their eyes and subsequently report hearing the *ba* syllable that is actually coming out of the speakers, that is. At which point most subjects will refuse to believe that what they hear with eyes closed is really the same audio feed as they heard when they watched the video, so striking is the illusion.

> module is informationally encapsulated: the only kind of information that can influence it is its designated input level. Through the chaining of integrative and inferential processors – and the possibilities for constrained feedback among them – we achieve overwhelmingly complex mapping between acoustic information and meaning. Furthermore, if each processor is mandatory and fast, then the chain will be mandatory and (almost as) fast. That is, the effect of Fodor's faculty-sized module is created by the chaining of a series of structure-specific modules (Jackendoff, 2002: 219-220).

Jackendoff's account of the structure-constrained modularity of language perception dovetails nicely with the proposal of Carruthers regarding how holistic reasoning could arise as a "virtual faculty" from the chaining and integration of a number of other, lower level, modular components. This explains the distinctly human ability to engage in holistic global reasoning without insisting that such reasoning consists of a *faculty-wide* module. Fodor's main objection to massive modularity was that such holistic reasoning couldn't be modular because local encapsulation would make the necessary epistemic background inaccessible to revision and, thus, global processing unfeasible. But this objection only makes sense if abductive inference, for example, were viewed as a single module in and of itself, subject to some extrinsic boundaries of its own encapsulation. Jackendoff suggests "that we can't talk properly about *the* informational encapsulation of a module. Rather, we have to talk about the informational encapsulation in relative terms" (Jackendoff, 2002: 228).

The point to take away from this discussion of Jackendoff is that while Carruthers' vision of "wide-scope encapsulation" may indeed turn out to be problematic, it is *not* so in a way that leads back to a Fodorian objector's position. Rather, the "virtual faculty of abduction" that Carruthers posits is encapsulated at numerous individual levels (in terms of the relative encapsulation of its component modules) and hence can be looked at as *modular* in exactly the same way that language perception is: the *appearance* of a faculty-wide, encapsulated module which is merely the chaining together and integration of a number of lower level domain-specific encapsulated modules. As such, the objection that Carruthers has abused the notion of modularity to claim it for his account does not seem to have too much bite. The "wide-scope encapsulation" he talks about might be better explained as a *virtual* encapsulation – since an integrative faculty will be exactly as encapsulated as the sum of its constituent parts are encapsulated relative to one another. Carruthers has not loosened the concept of modularity to the point of losing its meaning, although he may be guilty of muddying the waters with the infelicitous use of the oxymoronic sounding "wide-scope encapsulation" to apply to a *virtual* faculty. According to Jackendoff's formulation, a virtual integrative faculty can actually be in essence more *narrowly* encapsulated than its component modules, as interface modules can allow for even smaller subsets or bottlenecks of information to get through the process. In this way, Carruthers can explain very well how quick, computationally tractable abductive inference can take place, as the outputs of numerous individually encapsulated processes can be brought together through interfaces and sorted into *very* narrow bands of possible response, which could be processed quickly. Therefore, the complaint that Carruthers' concept of modularity is too wide in scope to benefit from the sort of computational frugality engendered by 'modularity' in the Fodorian sense, is actually off the mark. As Jackendoff's model shows, a module made up of several sub-component modules actually may end up working from an even *smaller* database than any of its lower level, upstream components. As the domain may widen in one sense (i.e., the *types* of inputs made available), the focus *within* the wider domain is narrowed vi bottlenecks, and the actual computational needs of the over-arching 'module' are made even simpler.

To come at this question from a different direction, it is useful to look at another human cognitive capacity that is often suggested as a clear candidate for modularity: the folk psychology or theory of mind faculty.[32] The ability to ascribe intentional and epistemic states to other people is one of the hallmarks of human cognition – and one which appears to meet many of the Fodorian criteria for modularity, as it is (at least) reflexive, mandatory, quick, developed along a characteristic ontogenetic course, and subject to systematic breakdown.[33] Simon Baron-Cohen presents an elegant modular account of a theory of mind faculty comprised of three native subcomponent modules, which develop according to specific schedule and build upon one another to create a fourth (one might say 'virtual') theory of mind module.

The first module to develop is what Baron-Cohen calls the *Intentionality Detector (ID)* which "works through the senses (vision, touch, and audition), and its value lies in its generality of application: it will interpret almost anything with self-propelled motion, or anything that makes a non-random sound, as a query agent with goals and desires" (Baron-Cohen, 1996: 34). This *ID* fits all the criteria for modularity: it is universal, reflex-like, automatic, fast, and prone to "illusions" or breakdowns where non-intentional objects are anthropomorphized and ascribed intentions if they move as if under their own control.[34] After the *ID* comes online, a normal-developing infant will begin to show signs of an *Eye Detection Detector (EDD)*, which has only "three basic functions: it detects the presence of eyes or eye-like stimuli, it computes whether eyes are directed towards it or something else, and it infers from its own case that if another organism's eyes are directed at something then that

[32] Keeping in mind that Carruthers' account relies on there being such a module of folk psychology, as outlined in chapter two.

[33] It would be beyond the scope of this thesis to really examine those claims in detail, and Fodor, for example, may not agree with such modular claims for our theory of mind faculty. See Baron-Cohen (1996), Baron-Cohen, Frith & Leslie (1995) Segal (1996), Andrews (2005), and Siegal & Surian (2006) for more discussion of the possible modularity of theory of mind.

[34] This was demonstrated in a series of experiments by Heider & Simmel (1944) using cartoon shapes that interact onscreen in such a way as to incite observers to impute goal-directed agency to them.

organism sees that thing" (Baron-Cohen, 1996: 38-39). In this case it seems even less difficult to grant such a processor the status of 'module', as it has an extremely specific domain: eyes or eye-like entities.[35]

According to Baron-Cohen, a third module, the *Shared-Attention Mechanism (SAM)*, develops to take inputs from both *ID* and *EDD* to create triadic representations of shared attention between the self and another agent.

> It then computes shared attention by comparing another agent's perceptual state with the self's current perceptual state. It is like a comparator, fusing dyadic representations about another's perceptual state and dyadic representations about the self's current perceptual state into a triadic representation. Doing this allows the SAM to compute that you and I are both seeing the same thing [...] (Baron-Cohen, 1996: 46).

Furthermore, SAM is capable of "making *ID's* output available to *EDD*. This allows *EDD* to read eye direction in terms of an agent's goals or desires" (Baron-Cohen, 1996: 48). Claims for modular status for *SAM* are, again, quite compelling, at least in terms of how it may be subject to systematic patterns of breakdown. It is Baron-Cohen's contention that "available evidence points to a massive impairment in the functioning of *SAM* in most children with autism. Children with autism often do not show any of the main forms of joint-attention behaviour" although they do exhibit behaviour consistent with having both *ID* and *EDD* (Baron-Cohen, 1996: 66). Interestingly, congenitally blind children *can* establish joint-attention despite obviously not having access to eye detection themselves (as *EDD* would not be getting any ocular input). Even more surprisingly, however, blind children appear to understand implicitly the nature of *another agent's sight*, as they can respond correctly to instructions such as "show Mommy the object" and "make it

[35] And again, is prone to misfires due to illusion, as 'eye-like entities' readily jump to our attention when they appear randomly in the physical world (e.g., in cloud formations, or the knots in a plank of wood). Similarly, the eyes of painted portraits may appear to be "following" us as we look at them from different vantage points.

so Mommy cannot see the object" (Baron-Cohen, 1996: 67). Blind children *appear* to have a working *EDD* online; they can't feed any direct input to it, but they can avail themselves of some of its representational power regarding shared attention. On the other hand, children with autism can perform tasks which separately suggest functioning *ID* and functioning *EDD*, but they cannot link those two mechanisms together to create joint-attention representations, leading to Baron-Cohen's belief that there must be a *SAM*, and that it works in blind children despite the *EDD* not getting any direct perceptual information, yet fails in children with autism.

The actual *Theory of Mind Mechanism (ToMM)* comes online last, according to Baron-Cohen, "triggered in development by taking triadic representations from *SAM* and converting them into M-representations," or representations of the epistemic states of other agents (Baron-Cohen, 1996: 55). *ToMM* is what allows us to ascribe belief to other organisms, and additionally, to understand the referential opacity of the epistemic states we ascribe to them (i.e., the notion that what the other may believe might indeed be incomplete or false).[36] *ToMM* also coincides with the development of the capacity in infants for *pretend play* as "the mental state 'pretend' is probably one of the first epistemic mental states that young children come to understand" (Baron-Cohen, 1996: 53).[37] From this initial experience of a personal epistemic state that is not the same as a physical state, (e.g., the banana can be a 'phone', but the banana is not, actually, a *phone)* children develop "an adult-like ontology dividing the universe into mental and physical entities. Thus they appreciate that a real

[36] In this sense, *ToMM* introduces dramatic irony into a person's worldview, as one can know something the other does not, and *know that the other does not know*. This opens the door to the possibility of deception, which plays such an important role in the discussions of chapter two above concerning "cheater detection." Further discussion of the relationship between *ToMM* and irony will occur in later sections of this chapter.

[37] Remember that pretension and its relation to a theory of mind module were key points in Carruthers' account of human reasoning, above.

biscuit can be seen by several people, can be touched, and can be eaten, whereas a thought-about or dreamed-about biscuit cannot" (Baron-Cohen, 1996: 54). From this point of development, all of the necessary ingredients are present for the complex theorizing we regularly do regarding the mental states of others and how they correspond to our own.

The point of engaging in this discussion of Baron-Cohen is not to demonstrate that the *ToMM* is modular, although Baron-Cohen clearly suggests that it should be viewed so. Rather, I highlight his account here to underscore the point that Jackendoff also makes, which is that mental faculties comprised of component modules linked via interface mechanisms can operate in ways that *appear* almost wholly unencapsulated, yet when these faculties are fractionated into their subcomponents, we see that each individual module is quite clearly domain-specific. *ID* and *EDD* both have narrowly restricted input domains, *SAM* can take only outputs from those two, and *ToMM* is triggered solely by the development of *SAM*[38] and the capacity for pretense. So here, again as in language perception according to Jackendoff above, we see a faculty-wide process that *appears* to access a great deal of information in order to do its work, yet that work is nonetheless fully computationally tractable, as the information has already been sifted to a great degree by lower level modules and interface bottlenecks. The original criticism of Carruthers in this section was that his account stretches what 'module' means to the point of being just "a box in an organizational flowchart" and thereby not able to claim modularity, along with the attendant computational benefits of encapsulation, for itself. The conclusion from these discussions of both Jackendoff and Baron-Cohen suggest that Carruthers' does have the right to refer to his over-arching 'module' of abductive reasoning, and, accordingly, that a

[38] This is why Baron-Cohen suggests that *ToMM* is directly tied to *SAM*, as autistic children who (on his account) have an absent or impaired *SAM*, are found to have a similarly impaired *ToMM*, leading them to suffer from "mindblindness."

Fodorian objector should not be able to get very far with a critique along these lines. Carruthers should amend his account and drop the term 'wide-scope encapsulation' which seems to be the problem, in its oxymoronic construction. The suggested term 'virtual encapsulation' is likely less vulnerable to criticism and closer to the heart of Carruthers' argument.

3.2 *Empirical predictions*

A second potential weakness in Carruthers' account is that it lends itself to a number of empirical, testable predictions about human cognition and behavior, which if they turned out to be untrue, might cast doubt on the entire project. Carruthers points to one explicitly, regarding the mental rehearsal of action-schemata, which forms the basis of his account of human creativity.

> These ideas lead to a straightforward empirical prediction (but not necessarily one that is easy to test). I postulate that no one could be a composer who was neither capable of playing an instrument nor of singing (nor perhaps of writing music, to rule out the possibility that the creative process involves *visual* images of notes on a stave, created by manipulations of motor images of the writing process). I know of no counter-example to this hypothesis (Carruthers, 2006: 310).

The idea here is that creativity comes from cycles of mentally rehearsed action schemata, so it follows that if one *lacks* a certain type of action schemata, one could not be creative in that context.[39] Carruthers is quite sure that this is the case, although it actually seems a little dubious. His prediction implies that the composer *hears* the music in the mind only as a result of mentally activating stored 'play' or 'sing' schemata and reassembling those. But it seems entirely plausible, and somehow more likely, that a composer *literally hears* music in the

[39] Which is *not* to say that a musician couldn't play a totally novel set of notes, since the mathematical 'grammar' of music allows for infinite recombination, iteration, and extrapolation of the musical alphabet. The point is that without that "grammar" (without having played *some* notes in the past), one could not creatively generate any action schemata to "play" in the mind in order to compose or create new music.

head, perhaps broadcasting memory outputs back through the auditory processing system, and some of the *mathematical* modules which surely play a role in comprehending *relative* pitch (which is the basis of harmony and melody – the judgment of *intervals* of sound). It seems that anyone who has ever *heard* music (or anything, really, since all sound has the necessary elements of 'music': aural frequency, rhythm, timbre, duration, amplitude, etc.) – anyone with hearing – could be capable of composing music *at least in their head*.[40] The question of how they would then *express* that composition is a different one – as they would have to learn to play an instrument or sing or notate music in order to express the composition, but this is not to say that the composition itself was inconceivable *before* the ability to express it. For this reason, I believe that Carruthers' empirical prediction is quite wrong. The larger question however, is if this actually poses a fatal problem for his greater account, or whether he is simply jumping the gun and making an empirical prediction that his own theory does not actually support.

Baron-Cohen's work again may be helpful regarding this question. As we have seen in section 3.1 above, congenitally blind children can nevertheless make sense of propositions involving the sightlines of other people (in terms of presenting and/or hiding objects from the sight of another). This seems to be a clear case in which a person is entertaining action schemata which s/he has no direct experience of him or herself. If some alien organism suddenly approached me and tried to explain some form of extrasensory perception it was capable of, I would likely be completely unable to make sense of it, or to readily make mental

[40] Beethoven will immediately come to mind here, as he was famously deaf at the end of his career, composing the Ninth Symphony while totally unable to hear it performed. Of course, Beethoven *had* his hearing previously, so his example does not serve much use here. The question of whether a congenitally deaf person could be a composer is, however, an interesting one. I would argue in the affirmative, for the reasons just outlined above: that music is essentially a mathematical enterprise, with a mathematical 'language' that could arguably be learned and understood perfectly well even by someone incapable of perceiving the sound that is produced, or able to experience the qualia of aesthetic aural perception of music, much in the same way the congenitally blind intuit sightlines.

representations of what the experience is *like*. Yet, the congenitally blind child seems thoroughly capable of understanding the *principle* of sight perception, despite having no actual experience of it. In a sense, the blind child can rehearse the action schema of seeing (in order to mentally model the process as it occurs in another person), without ever having engaged that action schema directly him or herself. This is likely because there are other underlying modular mechanisms involved in the overarching faculty of sight downstream from the initial ocular input system, which conceivably are present in the blind *despite* the absence of ocular input. However, in my conversation with the alien, I lack not only the initial perceptual inputs, but also the equivalent downstream processors for this alien sixth sense, and hence cannot reason sensibly about what the peculiar sensory experience consists of.

Carruthers' prediction regarding non-musicians' inability to compose seems to treat the composer as analogous to the alien in the above distinction, as he is arguing that one must have prior experience with a particular action schema to use it in creative ways. However, it seems as if the non-musician could be deemed more analogous to the blind child, who *can* still access the action schemata involved in sight, just from a more indirect route. Similarly, perhaps the non-musician could be able to engage action schema regarding musicality or composition via some indirect route. Therefore, it seems unlikely that Carruthers' empirical prediction is well founded. The next question is how deep a problem this predictive failure may be for his account.

One empirical finding that may support Carruthers' empirical prediction *in principle* (although not in the *specific* case of music composition) is research involving pattern recognition and memory in master-level chess players. When shown images of chess boards in various states of play and then asked to reconstruct the positions of the pieces from memory, chess masters perform demonstrably better when the image they are shown is one depicting the piece positions of an actual game in progress than when the image is one of just a random distribution of pieces on the board (Chase & Simon, 1973). This suggests that board positions reachable through actual play are more easily processed in memory than positions that are not the result of actual play. Yet this finding seems to suggest an impossible computational task: namely, the ability to quickly reverse-engineer a board position to determine whether or not it is likely the result of actual play and therefore worth bringing one's mental chess abilities to bear upon it. Surely this is not what is happening, as the exponential volume of permutations of possible moves involved would take far too long to mentally run in reverse. The answer obviously lies in some sort of pattern recognition process, whereby chess positions that are the result of actual play trigger memories of past games played in which similar positions were achieved, hence engaging the 'chess-brain,' as it were, to allocate more resources to memory in those instances, thereby conserving computational resources (heuristically framing some boards as being within the domain of chess expertise and others not). The random board positions would likely not resemble any 'known' chess positions, and therefore would not kick the chess brain into high gear to facilitate better memory storage in those cases. The amateur chess player, on the other hand, should be unable to perform any better in reconstructing the board positions of actual games versus random distributions, which is exactly what the research bears out.[41]

These findings do seem to lend support to Carruthers' claims about the centrality of mental rehearsal of action schemata in the mind, and about how that may be tied to faster and more effective memory retrieval, though they say nothing about Carruthers' exact empirical prediction regarding musical composition, as the case of the chess master is quite

[41] If the amateur *could* perform as well as the master in this task, it would imply, quite improbably, that there either exists some native database of possible chess play patterns in the mind, *prior* to actually having played a great deal of chess, or that we *do* in fact have remarkable reverse engineering skills when it comes to the permutational deconstruction of chess move histories.

differently formulated. The success of the chess master in being able to remember and reconstruct *playable* positions, however, is exactly the kind of result that Carruthers' account of cognitive architecture should be able to predict (had the experiment not preceded him, of course), as *playable* positions are only recognizable and mentally manipulable as such by *players*. In this sense, Carruthers' may be on the right track with his prediction, but has simply not chosen his example very well to illustrate what he is getting at. However, even if his specific prediction turns out to be unfounded, this fact shouldn't undercut Carruthers' underlying account, as other empirical findings *do* lend support to his model.

3.3 *The specter of circularity*

A final definitional cavil which could be made against Carruthers is potentially much more serious, having to do with the sometimes fast and loose way he throws around the terms 'context' and 'relevance' in his account of cognitive framing. Since the *mystery* of how relevance is ascertained is absolutely central to the Fodorian input problem, any model which hopes to overcome that objection needs to be quite circumspect in how 'relevance' is employed in the account. For example, Carruthers adopts a Sperberian 'satisficing' policy when it comes to using heuristics to ensure computational tractability in reasoning chores, but the *satisfaction* is always couched in terms of being satisfactory *in the 'given context'*. There is a potential element of circularity in this explanation.[42] What is it that causes the subject to be satisfied by the effect of the employed heuristic? Presumably the answer is that some 'desired outcome' has been achieved – one which satisfies the motivational system which set

[42] Using 'context' as part of any *explanation* can be dangerously circular, as Austin reminds us when it comes to language also: "for some years we have been realizing more and more clearly that the occasion of an utterance matters seriously, and that the words used are to some extent to be 'explained' by the 'context' in which they are designed to be or have actually been spoken [...] Yet still perhaps we are too prone to give these explanations in terms of 'the meanings of words'" (Austin, 1962: 100).

the action-schema in motion. But the desired outcome is obviously situation-dependent, which implies that *some sort* of judgment of the given context had to have been presupposed insofar as it triggered a cascading cycle of mental rehearsals of action schemata and inner speech (on Carruthers' account) which result in the action(s) which are now found to have led to the desired outcome. So, *yes*, Carruthers' architecture can handle the individual operations each step of the way without any sort of 'holistic' sense of relevance, but it does seem like there is some sort of holistic 'understanding' of the 'given context' and a holistic sort of understanding of the *desired goal state to be satisfied* in order to trigger the entire process. Carruthers' account postulated a "somasensory monitoring system" to do this job, but it is unclear how such a system could claim any sort of modular, computationally tractable status, as it would appear to have the entire mind within its purview.

I suppose Carruthers' answer to this would be that goal states are the result of interaction between desire- and belief-generating systems which have resulted in past action schemata and *memories* of the results. New situations, or contexts, can trigger the same motivational states as they are *associated* via memory priming effects to past events, and hence the 'context' is established. However, this seems to be putting a great deal of pressure on memory databases, which presumably grow to be quite deep over time, and again forces us back to wondering *how* memories are called up out of the depth of storage via feature association, because isn't *feature association* also a type of *decision* or judgment of similarity? Unless the argument is that there is some physical proximity of like memories in the physiological / neural structure of the brain that allows for such easy triggering,[43] the 'library' model of domain general memory which Carruthers has turned to in previous papers (though not overtly in the 2006 book) has the problem of knowing *where to look*. Is there a

[43] Which, once again, suggests some sort of unitary executive that can tag memories as similar and hence place them in the relevant cortical areas. The ghost keeps reappearing.

Dewey decimal system for memory storage? And if so, where is the homunculus who wrote it? S/he seems to be the kind of ghost that Fodor is always expecting to meet.

The questions of circular reasoning piling up here suggest a two-pronged critique of Carruthers: one regarding the nature of the somasensory monitoring system that would be required to contextualize cognition in terms of the (constantly shifting) goal/desire states of the individual, and a second regarding the ins and outs of memory priming and how that can be explained within a modular, computationally frugal system. We will look at each of these problems in turn.

First, regarding the notion of the somasensory monitoring system: would not the assumption of somasensory monitoring require some sort of holistic central executive that is capable of doing the monitoring? And if so, wouldn't this imply some sort of global outlook and a wide grasp of the background somasensory state, not to mention an ability to make objective judgments, or at least measurements concerning it? If human cognitive architecture is predicated on an underlying belief/desire psychology, then it seems as if any somasensory monitoring system would exist *within* that structure, and as such, be unable to make the sort of neutral, objective assessments that Carruthers' account seems to require it to make, in order to adjust levels of motivation, expectation and satisfaction on the fly.

In Sperber and Wilson's account, some similar sort of system is required to make *satisficing* determinations regarding the "cognitive utility" of processing any thought or piece of information. "*Ceteris paribus*, the greater the effect of processing a given piece of information, the greater its relevance. *Ceteris paribus*, the greater the effort involved in processing a given piece of information, the lower its relevance" (Sperber & Wilson, 1996: 531). Yet, how are such cost/benefit analyses supposed to take place? The idea that relevance can be framed as a result of a cost/benefit analysis of the processing required seems to imply a global executive monitoring system that can make such objective judgments. Without that, then we may be forced into suggesting that such a monitoring system *itself* relies on satisficing principles, and fall sideways into the type of regress Fodor warns us about.

Sperber and Wilson resort to a physiological analogy in order to try and explain how it might work:

> Here is one line of possible speculation: contextual effects and mental effort, just like bodily movements and muscular effort, must cause some symptomatic physico-chemical changes. We might assume that the mind assesses its own efforts and their effects by monitoring these changes (Sperber and Wilson, 1986: 130).

The analogy to how resources are allocated to muscles within the body does not seem too helpful in this case, because, although it is *true* that different bodily systems ramp up or tone down activities according to physico-chemical changes, in the case of bodily systems, this is precisely *because there is a mind/brain to coordinate the activity.* The human body *has* a global executive function (the mind/brain) whose domain is *the entire body.* The mind/brain can coordinate signals from various bodily systems, and allocate resources as needed (by sending physico-chemical messages to appropriate systems). To extend the analogy to how the mind frames context in the world still leaves the specter of a *mind within the mind,* a global executive function that has the entire mind (and world, including the minds of others) as its domain, and can calculate cost-benefit analyses and compute both cognitive effort and effect based on a clear picture of the current somasensory state of the organism.

Carston perhaps offers a more charitable and effective interpretation of what Sperber and Wilson seem to be arguing here:

> A basic principle of the framework is the *Cognitive Principle of Relevance* according to which the human cognitive system as a whole is oriented towards the maximization of relevance. That is, the various subsystems involved, in effect, conspire together in a bid to achieve the greatest number of cognitive effects for the least processing overall. The perceptual input systems have evolved in such a way that they generally

> respond automatically to stimuli which are very likely to have cognitive effects, quickly converting them into the sort of representational formats that are appropriate inputs to the conceptual inferential systems; these systems then integrate them, as efficiently as possible, with some accessible subset of existing representations to achieve as many cognitive effects as possible (Carston, 2002: 45).

The main point of contention here would still be the integration phase, for it seems to imply a fairly global outlook in terms of routing information from inferential systems to appropriate subsets of representations. This will inevitably lead to a discussion of memory priming, which will covered in more depth below. However, this is an infinitely more satisfactory explanation that that offered by Sperber and Wilson, who simply suggest "the mind assesses its own efforts" at computing relevance, which is hopelessly circular.

Levinson is highly critical of Sperber and Wilson on this point, noting "the factor of cognitive effort, an essential ingredient in the proportional measurement of relevance, is not empirically measurable (or at least not empirically measured)" (Levinson, 2000: 57). This point is well taken: whatever principle of relevance the mind employs in framing cognitive tasks is not one that submits itself to empirical measurement. And yet, it *is* employed – we do mentally assign or infer relevance and context – the question is how we do so in a computationally tractable way when it appears to be such a global, computationally explosive process. So from where is the principle of relevance employed? Is there another alternative than that it is employed by a ghost-like domain-general processor within the mind? We can turn briefly to a discussion of context framing and relevance in linguistic communication in order to seek a clearer answer.

Sperber and Wilson originally brought their principle of relevance into the cognitive realm from the *communicative* realm, where context and relevance play key roles in the linguistics discipline of *pragmatics*, which is the study of how speaker intention is ascertained in communication. We impute speaker intention and frame contexts in communication quite efficiently, against quite formidable computational odds, and from a very young age; so if pragmatics has an answer for how this process works in communication, we should, in principle, be able to extend that answer back to cognition, hopefully in a manner more coherent than that of Sperber and Wilson. Grice suggests that this should be an achievable goal, as "use of language is one among a range of forms of rational activity and those rational activities which do not involve the use of language are in various ways importantly parallel to those which do" (Grice, 1989: 341). This strategy should be especially amenable to Carruthers, as his account highlights the centrality of language usage to human cognition, and holistic reasoning in particular. If context framing within language can be explained, then cognition via language in humans should be able to avail itself of the same processes.

So how much does pragmatics actually have to say about this subject? In the relevance-theoretic tradition starting with Grice, the assumption has been that the communicative act takes place within certain parameters which assign or presume relevance in terms of speaker intention.[44] What a speaker *means* by a given set of words is what the speaker *intends the hearer to understand.* This seems obvious enough, but when we actually start to look at language use, we find a whole slew of utterances which seem quite complicated in this regard. Many of the things we routinely say are ***underdetermined***, and require non-demonstrative inference in order to ascertain speaker intention. Utterances may be metaphorical, ironic, or contain implicatures which demand a full knowledge of the relative "Background" (to steal Searle's term) of the utterance in order to parse them correctly.

> How does one recognize another individual's intentions? [...] the problem is not that it is hard to come up with hypotheses about what the communicator might have intended to convey: it is that too many hypotheses are possible. Even a linguistic

[44] Arguably, this linking of meaning with intention or *use* could be laid at the feet of Wittgenstein, who suggests famously in the *Philosophical Investigations* that "for a large class of cases – though not for all – in which we employ the word 'meaning' it can be defined thus: the meaning of a word is its use in the language" (*PI*, § 43).

> utterance is generally full of semantic ambiguities and referential ambivalences, and is open to a wide range of figurative interpretations. For non-coded behaviour there is, by definition, no predetermined range of information it might be used to communicate. The problem then is to choose the right hypothesis from an indefinite range of possible hypotheses (Sperber & Wilson, 1986: 32-33).

Sperber and Wilson recognize the problem this poses, as "any conceptually represented information available to the addressee can be used as a premise in this inference process. In other words [...] this process is 'global' rather than 'local': where a local process is either context-free or sensitive only to contextual information from some set domain" (Sperber & Wilson, 1986: 65).[45] No logical deductive framework can explain how this inference takes place, and the only alternative is some form of "constrained guesswork", assisted by various heuristics (Sperber & Wilson, 1986: 69).

Grice formalized the heuristic constraints of communicative inference into nine so-called conversational "maxims"[46] which collectively form the "co-operative principle", working from the assumption that communicative acts "are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction [...] at each stage, *some* possible conversational moves would be excluded as conversationally unsuitable" (Grice, 1975: 45). In this sense, all communication takes for granted a degree of "co-operation" insofar as all parties to that communication assume a stance relative to one another that their utterances are meant to convey a specific meaning to the hearer, and assume an underlying commitment to cooperate in making that happen. In this way, inferences can be made on the fly as conversations progress, if all parties can assume their counterparts to be adhering faithfully to the same set of maxims. As Sperber and Wilson

[45] Interestingly, in their 1986 book, this leads Sperber and Wilson to *agree* with Fodor that inferential processing "is not a separate, purpose-built ability" and therefore must take place in a central, global system, and not a local or modularized system (Sperber & Wilson, 1986: 67).

[46] I won't take the time here to list all the maxims.

explain, the fundamental point Grice has hit on here is that "the very act of communicating creates expectations which it then exploits" (Sperber & Wilson, 1986: 37). Or as Sperber and Wilson put it in their own terms, "this amounts to saying that an ostensive communicator necessarily communicates that the stimulus she uses is relevant to the audience. In other words, an act of ostensive-inferential communication automatically communicates a *presumption of relevance"* (Sperber & Wilson, 1986: 156). Furthermore, this presumption need not be conscious, in the sense that Grice's is.

> Grice's principle and maxims are norms which communicators and audience must know in order to communicate adequately. Communicators generally keep to the norms, but may also violate them to achieve particular effects; and the audience uses its knowledge of the norms in interpreting communicative behaviour [...] The principle of relevance, by contrast, is a generalization about ostensive-inferential communication. Communicators and audience need no more know the principle of relevance to communicate than they need to know the principles of genetics to reproduce (Sperber & Wilson, 1986: 162).

It is not particularly clear how these "presumptions" are really getting us any closer to *how* we ascertain speaker intention in underdetermined utterances, however. The simple fact that all parties to a communicative exchange are agreed that speaker intention is important and are cooperating together to succeed in establishing that intention clearly (whether consciously or not) does not seem to explain very much at all. Certainly, it is better than some alternate situation in which parties to communication did *not* make such an assumption (in which case conversations would be aimless and random, one would think), but it still leaves *serious* computational challenges on the table, which will require specific heuristics to constrain, or a clear explanation of how we ascribe communicative intent to the utterances of others. Ultimately, we resort to a certain degree of *mind-reading* in order to do so – and this process apparently goes on mostly beneath the surface, and is not the result of deductive practices or linearly processed clarifications as conversations progress.

Givón, in *Context as Other Minds,* makes the argument that context framing in communication is exclusively a matter of employing our folk psychology or theory of mind faculty, insofar as

> [t]he mental construal of the mind of the other has, of course, been implicit in all works on communicative pragmatics. Even the most logic-bound treatments of 'intentional logic' (Carnap, 1956), 'definite description' (Geach, 1962; Strawson, 1964; Donellan, 1966), 'presupposition' (Keenan, 1969, 1972; Gazdar, 1979), 'conversational implicatures' (Grice 1968/1975; Levinson, 1983), or 'presumptive meaning' (Levinson, 2000) are suffused with assumptions about the mind of the other (Givón, 2005: 7).

To ascertain another speaker's intention is, according to Givón, to theorize about the epistemic and goal states of the other. On this account, our *ToMM* would need to have evolved *before* complex communication involving nondemonstrative inference could take place.[47] Givón explains why this is the most likely evolutionary order:

> Our innate heuristics and biases, not only those that involve 'primitive' 'affective' domains but also those that pertain to the most sophisticated higher-end cognitive strategies, did not evolve in a society where conversation with strangers was the most basic human ritual. Rather, they evolved, over 6 million years of separate hominid-line evolution, in the exclusive confines of a *society of intimates*, where all communication and cooperation took place among members of the intimate, small, kin-related group, and where interaction with strangers was overwhelmingly hostile [...] one communicated *only* with people one knew intimately and exhaustively, people about whom one could use, reliably, the tried and true calculus of *reasoning by feature association:* if you knew one trait of the person – their identity and kin relation to you – the rest of their traits, in particular their behaviour toward you in most relevant social contexts, was highly predictable (Givón, 2005: 223).

There are a number of important insights in this argument. First, if communication originally evolved in a "society of intimates" as Givón suggests, then it would make sense that our communicative abilities evolved to be able to exploit what Fodor would call the "very subtle clues" of speaker intention when communicating, as conversational partners were so exhaustively intimate to one another that they could *know what the other was thinking,*

[47] A claim which may well be supported by evidence showing that great apes, for example, can exhibit rudimentary theory of mind abilities, without having language. It would also go a long way toward explaining the dramatic communicative deficits in children with autism, if we are to agree with Baron-Cohen that they have impaired *ToMM.*

in the way only intimates can. Additionally, intimates ***know*** how well the other knows them, and therefore can subsume some of their intended meaning into such "very subtle clues" and expect that successful communication still will take place. As our social evolution led to greater interaction with strangers, a lot of these "very subtle clue" detectors would have been hardwired in place via natural selection as part of our *ToMM*, to the point that successful ostensive-inferential communication could take place quickly and efficiently, so long as the cultural differences are not too vast, or that there is no impairment to the *ToMM* of any party to a conversation.[48]

The other important insight of Givón's passage above is the idea that in framing communicative context we utilize "the tried and true calculus of *reasoning by feature association.*" We impute intention to another based, at least in part, on a comparison of the presumed context of their utterances to our own perceived intention. At first blush, this may seem as question-begging as the earlier discussion regarding somasensory monitoring (insofar as one could ask *how we know our own minds* in order to make the comparison). However, Givón is hinting a deeper point here, namely that "reasoning by feature association" can go both ways: we can know the mind of the other based on feature associations to our own, *but also vice versa.*

[48] The former should be obvious, as anyone who has traveled and interacted with distant cultures can attest. As practical real world example of the latter, I offer the following personal experience: working in a high school last year I had the experience of directing a play which included one mildly autistic child in the cast. In one conversational exchange I told him "Oh, by the way – there is an emergency rehearsal tomorrow after school to go over last minute script changes," to which he replied, "Ok." The next day, he failed to show up at the rehearsal, and when I confronted him later regarding his absence he said he didn't know that he was supposed to be there. "But we talked about it!" I replied, "I told you about the rehearsal and you said 'Ok'!" To which he replied that I "never told him *he had to be there.*" Of course, in my mind that is *exactly* what I had communicated to him, but he was unable to take my words as anything else than literal, totally blind to my state of mind and deaf to my implied meaning: he heard simply a factual statement that there was a rehearsal, and could ascribe no other meaning to the statement than that. I learned my lesson quickly, and the next time I needed his presence I asked for it explicitly.

> Once again, evolution appears to have conspired to frustrate the predilections of simple-minded reductionists, having created a classical complex, hybrid – muddled and impure but still eminently serviceable – adaptive compromise. That our representation of the mind of the other (3rd order) rebounds and eventually transforms our own self-representation (2nd order) into a 4th order construct is but an adaptive consequence of being a social, cooperative, communicating species (Givón, 2005: 236).

The idea here is that we engage in feedback loops of intention ascription both to the other and to ourselves, presumably in real time, as communication unfolds. "Feature associations" constantly prime, bolster, and refine our theories regarding the epistemic state of both the other and, in rebound, of oneself, and thus the context of any given communication emerges as it is, in essence, sifted out. This description of the process should sit very well with Carruthers, as it mirrors the way in which cycles of inner speech can result in emergent reasoning in his account.

Givón's proposal here also brings us neatly to the second potential area of circularity in Carruthers' argument that this section was dedicated to exploring, and perhaps already hints at a way out of the difficulty. The problem in question was regarding the process of *memory priming*, which plays such a crucial role in Carruthers' account, yet as discussed earlier, seems to imply some sort of executive filing system in order to tag *like* memories or information as *like* in the first place. Such a process certainly falls under the rubric of reasoning by feature association – the question is *how* are features associated? Judgments of 'similarity' seem to be essentially objective in a way that suggests the specter of a global executive has returned to the conversation.

One way out of this problem may lie in an analysis of *metaphor*. Indeed, metaphor, as a concept, appears to lay at a crucial intersection of the various threads being spooled out in this discussion: metaphor is a classic example of communicative implicature, metaphor relies on a working *ToMM* to be understood, and metaphor is predicated on feature association.

Lakoff and Johnson have a great deal of insightful commentary on the role of metaphor in both thought and language which may help bolster Carruthers' reliance on memory priming in context framing. Their primary contention is that "human thought processes are largely metaphorical [...] Metaphors as linguistic expressions are possible precisely because there are metaphors in a person's conceptual system" (Lakoff & Johnson, 1980: 6). They present voluminous evidence that many (most) of our basic concepts are metaphorically constructed and organized, commonly in spatial metaphors (e.g. IDEAS are OBJECTS, EXPRESSIONS are CONTAINERS, UP is POSITIVE, TIME is a JOURNEY, etc.)[49] Most of these metaphors are so internalized that we do not even recognize them as such (they are dead metaphors), but if so many concepts are organized around a relatively small set of basic spatial and orientational metaphors, then it becomes clear how we could have a simple system of memory priming available, without necessitating lengthy searches.

> [M]etaphors allow us to understand one domain of experience in terms of another. This suggests that understanding takes place in terms of entire domains of experience and not in terms of isolated concepts. The fact that we have been led to hypothesize metaphors like LOVE IS A JOURNEY, TIME IS MONEY, and ARGUMENT IS WAR suggests to us that the focus of definition is at the level of basic domains of experience like love, time, and argument. These experiences are then conceptualized and defined in terms of other basic domains of experience like journeys, money, and war. The definition of subconcepts, like ATTACKING A CLAIM and BUDGETING TIME, should fall out as consequences of defining the more general concepts in metaphorical terms.
>
> This raises a fundamental question: What constitutes a 'basic domain of experience'? Each such domain is a structured whole within our experience that is conceptualized as what we have called an *experiential gestalt*. Such gestalts are *experientially basic* because they characterize structured wholes within recurrent human experiences. They represent coherent organizations of our experiences in terms of natural dimensions (parts, stages, causes, etc.) Domains of experience that are organized as gestalts in terms of such natural dimensions seems to us to be *natural kinds of experience* (Lakoff & Johnson, 1980: 117).

[49] There is not the space here to really explain and give examples of the metaphorical bases of most of our concepts, for that you will have to refer to Lakoff and Johnson, 1980. Suffice it to say, the examples are ubiquitous.

All incoming information can get tagged via such conceptual metaphors to be associated to basic domains of natural kinds of human experience, hence creating a sort of virtual index for stored concepts that can be used to prompt associated concepts as needed. Givón explains how useful the idea of conceptual metaphors can be, inasmuch as they can effect "a *mapping* relation between *distinct cognitive domains.* Similarity, construed or otherwise, is presumably not involved [...] cross-domain relations are equations (X is Y) rather than similes (X is like Y)" (Givón, 2005: 75). This is key, as judgments of 'similarity' are what suggest some objective arbiter at the conceptual level. On the other hand, if the cross-domain relations are simply tagged as equations via conceptual metaphors, then there need not be any quasi-objective *measurement* taking place on a conceptual level – and hence no difficult computation to engage in. There is also added computational frugality in working with figurative images rather than strictly syntactic mentalese. As Paivio and Walsh note, "imagery ensures processing flexibility" compared to linear, sequential (verbal) processing.

> The synchronous nature of imaginal representations promotes efficient memory search because such information can be processed in a way that is flexible and relatively free from sequential constraints. If you are asked how many windows there are in your house, you can arrive at the answer by imagining your house from different angles and counting the windows from the image [...] in either direction, inside or out. By contrast, the processing of organized verbal information in long-term memory is sequentially constrained to a high degree. We can recite the alphabet forward more quickly than backward, and backward recitation of a poem would be painfully slow (Paivio & Walsh, 1979: 323).

Lakoff and Johnson's account of conceptual metaphor suggests a certain kinship with the Wittgensteinian notion of *family resemblances* among language games – "a complicated network on similarities overlapping and crisscrossing: sometimes overall similarities, sometimes similarities of detail" (*PI*, §66). It is especially interesting to note that the metaphor of family resemblance implies an evolution over time – different language games (and different concepts) evolve from one another in such a way that certain elements are retained while others are lost and new ones are incorporated, but connections are still viable and directly relatable to experience – naturalized, in this sense. Furthermore, there is a direct Wittgensteinian connection to the issue of context framing, insofar as he contends that all communication rests on *agreement*, which means nothing if not the shared understanding of intention in communication. "If language is to be a means of communication there must be agreement not only in definitions but also (queer as it may sound) in judgments" *(PI,* §242). Now, a voluminous amount of ink has been spilt trying to understand whether this remark means agreement among *instances* of judgment over time, or agreement among *individuals* in a community, a full exploration of which would be far beyond the scope of this thesis, but there is much evidence to support the latter interpretation, and, if it is a correct interpretation, it is useful for our discussion here.[50] Such a *conventionalist* account of meaning would buttress the picture being presented here of context framing as a fundamentally community-based endeavour, insofar as it depends on communication and the construal of other minds and intentionality in order to categorize and map concepts into broad domains

[50] Kripke is the standard-bearer for the "community" interpretation (Kripke, 1984), alongside Michael Dummett and, to an extent, Norman Malcolm, while Baker & Hacker, Simon Blackburn, and Colin McGinn all strenuously object to imputing such "conventionalist" views to the later Wittgenstein. For the clearest indication that Kripke's Wittgenstein is likely the closest to Wittgenstein's intended reading of the issue, refer to the following selection from the *Lectures on the Foundations of Mathematics:*

> Suppose we are in this room inventing arithmetic [...] We have invented multiplication up to 100; that is, we've written down things like 81x63 but have never yet written down things like 123x489. I say to [Lewy], 'You know what you've done so far. Now do the same sort of thing for these two numbers.' -- I assume he does what we usually do. This is an experiment – and one which we may later adopt as a calculation.
>
> What does that mean? Well, suppose that 90 percent do it all one way. I say, 'this is now going to be the right result.' The experiment was to show what the most natural way is – which way most of them go. Now everybody is taught to do it – and *now* there is a right and wrong. Before there was not.
>
> It is like finding the best place to build a road across the moors. We may first send people across, and see which is the most natural way for them to go, and then build the road that way. Before the calculation was invented or the technique fixed, there was no right or wrong result (Wittgenstein, 1976: 95).

This passage seems to indisputably suggest Wittgenstein's "agreement" is meant to refer to agreement within a community, and not merely over instantiations of rule-following behaviour.

of feature association (or family resemblance). The Wittgensteinian notion of family resemblances is perhaps an even simpler way of explaining memory priming than Lakoff and Johnson's conceptual metaphor account, though Wittgenstein himself resorts to (another) metaphor to explain it: "we extend our concept[s] as in spinning a thread we twist fibre on fibre. And the strength of the thread does not reside in the fact that some one fibre runs through its whole length, but in the overlapping of many fibres" *(PI,* §§67).[51]

The issue of *agreement,* and the thread metaphor of Wittgenstein, leads back once again to Jackendoff, who argues something similar, although using a slightly different explanatory metaphor, in his account of how associative memory priming functions. In *Language, Consciousness, Culture,* Jackendoff sketches an account of linguistic meaning construction via *parallel architecture* in which

> the processor does not arbitrarily choose among the possibilities and then go on from there (algorithmically). Rather, it constructs *all* reasonable possibilities and runs them in parallel, eventually selecting a single most plausible or most stable structure as more constraints become available, inhibiting other structures [...] I find it useful to think of the process of construction as achieving a 'resonance' among the linked structures, a state of global optimal stability within and among the structures in a complex. Occasionally among the promiscuous structures there are multiple stable

[51] Wittgenstein, despite writing long before and far outside the specific scope of this discussion, has a number of strikingly insightful comments to make on the issue of context framing and relevance determinations, although they could only be applied somewhat elliptically and do not offer answers to the *how* questions, but merely point to the centrality of conceptual frames and their effect on the tractability of thought. Wittgenstein notes that all thought is framed "like a pair of glasses on our nose through which we see whatever we look at" *(PI,* §103) ultimately concluding that "a *picture* [holds] us captive. And we [can] not get outside it, for it [lies] in our language and language [seems] to repeat it to us inexorably" (*PI,* §115). As for our endeavour here to try and explain how those frames operate, and how we impose them, Wittgenstein would likely dismiss it as pointless, suggesting that this is "simply what we *do"* and that our attempts to explain the phenomenon will fail; our "spade will be turned", "all our explanations [will] come to an end" and we will just have to accept that framing is part of our "form of life". (If nothing else, Wittgenstein knows how to stop thinking!) Nonetheless, I do think a highly interesting and novel paper could be written on "Wittgenstein the Evolutionary Psychologist," as so many of his remarks in the *Investigations* could be read through that lens in curiously insightful ways.

> states, in which case perception produces an ambiguous result such as the Necker cube in vision and a pun or other ambiguity in language (Jackendoff, 2007: 20).[52]

Memory priming is a key element of this, although one can see right away how this 'parallel' processing architecture could make memory searches easier. In essence, any given concept will 'light up' all other memories that may be related, however tangentially, *regardless* of the current context. All of these memories would then be made available to interface and integrative modules which would run them through, perhaps repeatedly, strengthening some structures which seem more stable given the ongoing flow of communication, and inhibiting dead ends. Jackendoff offers some empirical evidence of what he calls this "semantic promiscuity":

> [I]t is found that, for a brief period, the word *bug* heard in any sentential context primes (speeds up reaction time to) the 'lexical decision task' of recognizing either *insect* or *spy* as a word; these words are semantically related to different senses of *bug*. After this brief period, only one of those words continues to be primed: the one related to the sense of *bug* in the presented sentence (Jackendoff, 2002: 209).

Jackendoff additionally cites Bock and Loebell's (1990) research, showing that "not only do words prime other words, but syntactic structures prime other syntactic structures" (Jackendoff, 2002: 217). The key idea here is that priming is *not* a linear operation in which an individual mental representation is triggered by another as a result of being somehow contextually related to it (which seems to imply a global sorting function, or a pre-awareness of the prevailing context). Rather, the promiscuity theory holds that *numerous* representations are constantly being cycled up from long-term memory to working memory on the basis of semantic or syntactic connections, no matter how strong or weak, and *regardless* of context. Context will whittle them down as the process goes on, but only as a result of cycling primed

[52] Kent Bach argues something similar in terms of how we understand implicatures, that we cycle through various inferential interpretations and settle on the most plausible – a sort of abductive process of meaning construction (Bach, 1999).

representations through various interface and integrative modules and, roughly speaking, seeing what sticks (or what achieves "resonance" in Jackendoff's terms above).

This may sound similar to Baars' *global workspace* account, or his blackboard analogy (discussed in chapter two). Jackendoff prefers a "workbench" analogy, but in general, it works the same way: if various modular processors post up all their outputs in a global workspace so that they may be grabbed as inputs by any other processors to whom they are relevant, then this could be another example of a *seemingly* executive function that is in fact not nearly as global as it appears, since it is no problem for modules to *recognize* inputs appropriate to them (especially given the enzymatic model of Barrett discussed in chapter two). The relevance of input to module is in this case tagged by selection, not a deliberative process – and hence there is no danger of slipping in a global executive. As Jackendoff concludes

> I want to think of working memory not just as a shelf where the brain stores material, but as a workbench where processing goes on, where structures are constructed. There seems no point in relegating processing to a 'central executive' when it has become abundantly clear that the brain is thoroughly decentralized (Jackendoff, 2002: 207).

It should be noted here that although Carruthers clearly aligns his "global broadcasting" picture with Baars "global workspace" account, Jackendoff is loathe to equate his "workbench" account with any type of global *broadcasting*, as Carruthers refers to. Jackendoff argues that the notion of broadcasting "cannot be sustained. A phonological structure, for example, is intelligible only to the part of the mind/brain that processes phonological structure. If that part of the mind 'broadcast' its contents to, say, a visual processor, it would be less than useless. And the same is true for any level of structure" (Jackendoff, 2007: 23). What are needed, according to Jackendoff, are interface modules which are bi-domain specific and can produce outputs which could then be taken up by different levels of structure. The Carrutherian global broadcasting picture oversimplifies how this would work, and this may be another source of the concerns outlined earlier that Carruthers is too loose in his formulation of modular processing and encapsulation: global broadcasting implies that most modules or levels of structure are capable of taking multiple input formats, which goes against the grain of encapsulation. Jackendoff's insistence that interface modules mediate such broadcasting is well taken, and serves to defend Carruthers' greater model from such critique.

Returning specifically to how all of this could help Carruthers, in should be clear at this point how associative memory priming could take place in practice without entailing a global executive tagging and retrieval mechanism. In essence, there *is* a 'Dewey Decimal System' of memory storage, however, it is not written by some ghost in the machine, but is rather inscribed on concepts via evolution within specific domains of natural experience, and is constantly being refined on the workbench of working memory as primed information is called up and processed. Carruthers' account perhaps did not include a clear enough explanation of how memory priming could work in this way, though it appears that memory can do what his proposed architecture needs it to do in order to function as he claims. We may now be in a position to close the book on the concerns that Carruthers may have skirted with circularity in his account of how we frame cognitive contexts. Indeed, we may be able to close the book on this entire chapter-long critique of Carruthers' account. It appears that whatever weaknesses his argument displays can be remedied by other accounts in such a way that his larger point stands fairly uncontested: there is a way to meet the challenge of the frame problem within a massively modular cognitive architecture. Furthermore, Fodor's input problem can be rejected, as it is possible to model and explain how abductive inference could take place on a purely local computational level, *without* entailing combinatorial explosion.

CONCLUDING REMARKS:
How to Stop Thinking

The aim of this thesis was to tackle the *frame problem*, and try to explain how human minds perform the seemingly impossible computational trick of framing relevance and context in communication and thought. All attempts to *build* an 'intelligent' machine have failed as a result of the recalcitrance of this problem regarding how to rule out the nearly infinite amount of information that *could* be brought to bear on any given problem and allocate cognitive resources where they will be most effective. Despite the prodigious processing power that machines are capable of compared to the relatively meager human mind, we don't have a clue, so far, how to make a machine mimic even the simplest contextual reasoning that a toddler is capable of. Human minds understand intuitively what is relevant in a given cognitive context, and therefore know what to focus on and what not to. Life can be extremely complicated, but it can also require quick decisions. We don't have the luxury of thinking through every possibility in a given situation before acting. We need to know how to *stop*. If we didn't have a way to cut off thinking and just *do something* every once in a while, we wouldn't last very long in a hostile world. Like Hamlet, or Dennett's fabled R-series robots discussed in the introduction to this paper, we would dither ourselves to death, quite literally. But we do not, or at least the psychologically healthy among us do not, suffer from such terminal indecision. We have evolved strategies which allow us stop thinking, arrive at decisions, frame contexts, and reach the point where, to paraphrase Wittgenstein, our cognitive spade is turned.

The central question of this thesis has been simply: *how do we do it?* The bulk of the discussion was predicated on an attempt to answer this problem from a reverse-engineering perspective – examining human cognitive framing in an attempt to determine how *exactly* it could work, without resorting to 'magical' answers, the *ghosts in the machine* that have haunted so many attempts to explain the human mind in the past. The primary assumption in this paper has been that there *is* a thoroughly materialist answer to how context framing is effectuated in the mind – ghosts are not the answer.

Chapter one was dedicated to outlining the frame problem in its various iterations, ultimately focusing on the version proposed by Fodor, who argues that cognitive science has so far rather ineptly bungled the question. The Massive Modularity thesis was put forward as one possible remedy, as encapsulated modular processes are by definition domain-specific, and therefore free of context effects. However, Fodor argues that such accounts are doomed *a priori* by what he calls 'the input problem', in which a modular mind courts a vicious regress as it requires mental representations to be routed to appropriate modules, but has no clear answer as to how that routing process can itself be modular. The first chapter laid out the stakes for modular theories of cognition: they must be able to explain the human phenomenon of abductive inference, or inference to the best explanation, which seems to imply a global computational process in which the entire epistemic background of the mind is part of the domain to be revised with each and every cognitive step. Fodor argues that massive modularity cannot account for abduction, and that there must be some sort of domain-general central executive processor, which we know next to nothing about, and which is essentially a ghost in the machine.

In chapter two, Peter Carruthers' account of a massively modular cognitive architecture was put forward as a serious contender to answer Fodor's challenge. Carruthers' argument was laid out in detail and an explanation was offered as to how, in principle, it could answer Fodor by positing a *virtual* faculty of abductive inference which emerges as an interaction effect between a handful of naturally selected modular processors, anchored by natural language, and facilitated by feedback loops of mental rehearsals of action schemata as 'inner speech.' In this sense, Fodor's abduction problem has been shown to be illusory, as abduction doesn't, strictly speaking, take place in any optimal sense; rather it is *virtual* – simply an *interaction effect.* Additionally, Carruthers' architecture of the mind was demonstrated to be capable of accounting for many of our distinctly human cognitive abilities, such as the generation of novel creative thought, and the ability to sensibly navigate the complex world of human sociality through communication by means of heuristically-enabled context framing and determinations of relevance.

A critique of three potential areas of weakness in Carruthers' account formed the basis of the final chapter of this thesis. The three soft spots identified in Carruthers' argument are 1) a terminological dispute regarding his usage of 'module' in application to faculty-wide cognitive functions, 2) a questionable empirical prediction that he suggests falls out of his account, and 3) some potentially circular reasoning as a result of an overly cavalier use of 'relevance' and 'context' in explaining relevance and context. The first and third concerns appear to be reconcilable when backed up by inquiries into psycholinguistics and pragmatics, whereas the second concern seems to be valid, although the fact that this one empirical prediction Carruthers makes turns out to be unfounded is not fatal to his account, but is merely the result of a misapplication of his own thinking. Jackendoff and Baron-Cohen offer many insights into the nature of modular cognitive functioning which support a Carrutherian view of over-arching, faculty-wide 'virtual modules' comprised of lower level component modules. Fodor's insistence that such a conception strips modularity of the encapsulation that defines it turns out to be overly dogmatic, and Carruthers can be said to be safely on solid ground here. As for the circularity of Carruthers' argument regarding context framing, he may indeed be guilty of some sloppiness, but a close reading of the pragmatic literature, married to the insights of Givón, and Jackendoff again, offer a way out for the Carrutherian model. Further tangential support for this bolstering of Carruthers, and some elliptical yet instructive insight into the nature of context framing can be found in Wittgenstein, which was noted in passing.

Ultimately, I believe that Fodor is mistaken: there is indeed an argument to be made that a massively modular cognitive architecture can adequately account for holistic common sense reasoning, and the appearance of abductive inference. It is possible to explain how we frame indefinite-seeming cognitive tasks in a computationally tractable way without resorting to demonic processors or ghosts in the machine. Carruthers' *Architecture of the Mind* largely succeeds in this effort, and where it falls short, like-minded accounts can be brought in to paper over the holes and preserve a convincing account. Certainly, a Fodorian objector will find numerous places to dig in to this thesis, and may well succeed in prying open more holes. However, I think the main *a priori* objections made to a modular account of context framing have been answered: a modular account *can* in principle answer the frame problem. The exact details of *how* will likely be a great source of future dispute, but the path ahead seems to clearly lie in that direction, Fodor notwithstanding. To end, we turn one final time to Wittgenstein, who seems to have seen us coming all along...

> Here we come up against a remarkable and characteristic phenomenon in philosophical investigation: the difficulty -- I might say -- is not that of finding the solution but rather that of recognizing as the solution something that looks as if it were only a preliminary to it. 'We have already said everything. -- Not anything that follows from this, no *this* itself is the solution!' This is connected, I believe, with our wrongly expecting an explanation, whereas the solution of the difficulty is a description, if we give it the right place in our considerations and do not try to get beyond it.
>
> The difficulty here is: ***to stop*** (Wittgenstein, 1967: §314).

REFERENCES:

Austin, J. L. (1962) *How to Do Things With Words.* Oxford: Oxford University Press

Baars, B.J. (1988) *A Cognitive Theory of Consciousness.* Cambridge: Cambridge University Press

Baron-Cohen, S. (1992) *Mindblindness* (Cambridge MA: MIT Press).

Baron-Cohen, S., Leslie, A., Frith, U. (1985) "Does the Autistic Child have a 'Theory of Mind'?" *Cognition* 21: pp. 37-46.

Bach, Kent. (1999) "The myth of conventional implicature." *Linguistics and Philosophy* 22: pp. 262-83

Baker, G.P., Hacker, P.M.S. (1990) "Malcolm on Language and Rules." *Philosophy 65,* pp. 167-179

Barrett, H. C. (2005) "Enzymatic computation and cognitive modularity." *Mind and Language 20, pp.* 259-287

Brown, P., and Levinson, S. (1978) *Politeness: Some Universals in Language Usage.* Cambridge: Cambridge University Press.

Carruthers, P., Smith, P. (1996) *Theories of Theories of Mind.* Cambridge: Cambridge University Press

Carruthers, P. (2000) *Phenomenal Consciousness.* Cambridge: Cambridge University Press

Carruthers, P. (2003) "On Fodor's Problem". *Mind and Language,* 18(5): pp. 502-523.

Carruthers, P. (2005a) "Distinctively Human Thinking" from *The Innate Mind* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Carruthers, P. (2005b) *Consciousness: Essays from a Higher Order Perspective.* Oxford: Clarendon Press

Carruthers, P. (2006). *The Architecture of the Mind.* Oxford: Clarendon Press

Carston, R. (2002) *Thoughts and Utterances: The Pragmatics of Explicit Communication.* Oxford: Blackwell

Chalmers, D. (1995) "Facing Up to the Problem of Consciousness" from *The Journal of Consciousness Studies,* 2(3), pp.200-219

Chase, W. G., Simon, H. A. (1973) "The mind's eye in chess" from *Visual information processing* (ed. W. G. Chase). New York: Academic Press, pp. 215–281.

Cheng, P. W., Holyoak, K. J. (1989) "On the natural selection of reasoning theories . *Cognition 33,* pp. 285-313.

Clark, A. (1991) "Systematicity, Structured Representations and Cognitive Architecture: A Reply to Fodor and Pylysyhn" from *Connectionism and the Philosophy of Mind* (eds. Horgan, T. & Tienson, J.) Dordrecht: Kluwer Academic Press, pp.198-217

Clarke, M. (2004). *Reconstructing Reason and Representation.* Cambridge: MIT Press

Chomsky, N. (1988) *Language and Problems of Knowledge.* Cambridge: MIT Press

Cosmides, L. (1989) "The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task" *Cognition 31,* pp. 187-276.

Cosmides, L. & Tooby, J. (1994). "Origins of Domain-Specificity: The Evolution of Functional Organization" from *Mapping the Mind: Domain-Specificity in Cognition and Culture* (eds. L. Hirschfeld, S. Gelman). New York: Cambridge University Press, pp. 85-116.

Dennett, D. (1987) "Cognitive Wheels: The Frame problem of AI" from *The Robot's Dilemma* (ed. Pylyshyn). Norwood, NJ: Ablex, pp. 41-64

Dummett, M. (1986) "Wittgenstein on Necessity: Some Reflections" from *The Seas of Language.* Oxford University Press: Oxford, pp. 446-461

Dunn, M. (1990) "The Frame Problem and Relevant Predication" from *Knowledge, Representation, and Defeasible Reasoning* (eds. H. Kyburg, R. Loui, G. Carlson). Dordrecht: Kluwer Academic Publishing, pp. 89-95

Festinger, L. (1957) *A Theory of Cognitive Dissonance.* Stanford: Stanford University Press

Fodor, J.A. (1983) *The Modularity of Mind.* Cambridge: MIT Press

Fodor, J.A. (1984) "Observation Reconsidered" *Philosophy of Science* 51(1), pp. 23-43

Fodor, J.A. (1987) "Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres", from *The Robot's Dilemma* (ed. Pylyshyn). Norwood, NJ: Ablex (pp.139-149)

Fodor, J.A. (1994) *A Theory of Content and Other Essays.* Cambridge: Bradford

Fodor, J.A. (2000). *The Mind Doesn't Work That Way.* Cambridge: MIT Press

Fodor, J.A. (2005) "Reply to Pinker's 'So How Does the Mind Work?'" *Mind & Language 20 (1),* pp. 25-32

Gazzaniga, M. (ed.) (1995) *The Cognitive Neurosciences.* Cambridge: MIT Press.

Geffner, H., Pearl, J. (1990) "A Framework for Dealing with Defaults" from *Knowledge, Representation, and Defeasible Reasoning* (eds. H. Kyburg, R. Loui, G. Carlson). Dordrecht: Kluwer, pp. 69-87

Gigerenzer, G. (2000) *Adaptive Thinking: Rationality in the Real World.* New York: Oxford U. Press

Gigerenzer, G., Todd, P. M. (1999) *Simple Heuristics that Make Us Smart.* New York: Oxford U. Press

Givón, T. (2005) *Context as Other Minds: The Pragmatics of Sociality, Cognition and Communication.* Amsterdam: John Benjamins Publishing Co.

Goodie, A., Ortmann, A., Davis, J.N., Bullock, S., Werner, G. (1999) "Demons vs. Heuristics in AI, Behavioural Ecology, and Economics" from *Simple Heuristics that Make Us Smart* (ed. Gigerenzer & Todd) New York: Oxford Univ. Press, pp. 327-356

Gopnik, A., Meltzoff, A. (1997) *Words, Thoughts and Theories.* Cambridge MA: MIT Press

Grice, H. P. (1957) "Meaning". *Philosophical Review*, 66: pp.377–88

Grice, H. P. (1989) *Studies in the Way of Words.* Cambridge: Harvard University Press

Gylmour, C. (1987) "Android Epistemology: Comments on Dennett's 'Cognitive Wheels'" from *The Robot's Dilemma* (ed. Pylyshyn). Norwood, NJ: Ablex, pp. 65-75

Hayes, P.J. (1987) "What the Frame Problem Is and Isn't" from *The Robot's Dilemma* (ed. Pylyshyn). Norwood, NJ: Ablex, pp. 123-138

Heider, F., Simmel M. (1944) "An experimental study of apparent behavior". *American Journal of Psychology 57,* pp. 243–259.

Jackendoff, R. (1992) *Languages of the Mind.* Cambridge: MIT Press

Jackendoff, R. (2002) *Foundations of Language: Brain, Meaning, Grammar, Evolution.* Oxford/New York: Oxford University Press

Jackendoff, R. (2007) *Language, Consciousness, Culture: Essays on Mental Structure (Jean Nicod Lectures).* Cambridge: MIT Press

Johnson-Laird, P.N. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness.* Cambridge: Cambridge University Press

Kripke, S. A. (1984) *Wittgenstein on Rules and Private Language.* Oxford: Harvard Press

Lakoff, G., Johnson, M. (1980) *Metaphors We Live By.* Chicago: University of Chicago Press

Lakoff, George. (1987) *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago: University of Chicago Press

Levinson, S. C. (1983) *Pragmatics.* Cambridge: Cambridge University Press

Levinson, S. C. (2000) *Presumptive Meanings: The Theory of Generalized Conversational Implicature.* Cambridge: MIT Press

Libet, B. (2004) *Mind time: The temporal factor in consciousness.* Cambridge: Harvard University Press

Lloyd, D. (1991) "Leaping to Conclusions: Connectionism, Consciousness, and the Computational Mind" from *Connectionism and the Philosophy of Mind* (ed. T. Horgan). Dordrecht: Kluwer

Malcolm, N. (1989) "Wittgenstein on Language and Rules". *Philosophy 64*, pp. 5-28

McCafferty, A. (1990) "Speaker Plans, Linguistic Contexts, and Indirect Speech Acts" from *Knowledge, Representation, and Defeasible Reasoning* (eds. H. Kyburg, R. Loui, G. Carlson). Dordrecht: Kluwer Academic Publishing, pp. 191-220

McGinn, C. (1984) *Wittgenstein on Meaning.* Basil Blackwell: Oxford

Nichols, S., Stich, S., Leslie, A., Klein, D. (1996) "Varieties of Off-line Simulation" from *Theories of Theories of Mind* (eds. Carruthers, P. & Smith, P.K.) Cambridge: Cambridge Press, pp. 39-73)

Ortony, A. (ed.) (1979) *Metaphor and Thought.* Cambridge: Cambridge University Press

Paivio, A., Walsh, M. (1979) "Psychological Processes in Metaphor Comprehension" from *Metaphor and Thought* (ed. Ortony) Cambridge: Cambridge University Press, pp. 307-328

Pinker, S. (1997) *How the Mind Works.* New York: W. W. Norton

Pinker, S. (2005) "So, How *Does* the Mind Work?" *Mind & Language* 20 (1): pp.1-24

Pylyshyn, Z.W. (1987) *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Norwood: Ablex

Samuels, R. (2005) "The Complexity of Cognition" from *The Innate Mind* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Schubert, L. (1990) "Monotonic Solution of the Frame Problem in the Situation Calculus" from *Knowledge, Representation, and Defeasible Reasoning* (eds. H. Kyburg, R. Loui, G. Carlson). Dordrecht: Kluwer Academic Publishing, pp. 23-68

Searle, John. (1979) 'Metaphor' from *Philosophy of Language* (ed. A.P. Martinich). Oxford: Oxford University Press

Segal, G. (1996) "The Modularity of Theory of Mind" from *Theories of Theories of Mind* (eds. Carruthers, P. & Smith, P.K.) Cambridge: Cambridge Univ. Press, pp. 141-157

Siegal, M., Surian, L. (2006) "Modularity in Language and Theory of Mind" from *The Innate Mind: vol.2* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Shakespeare, W. (1954) *Hamlet.* New Haven: Yale University Press

Shusterman, A., Spelke, E. (2005) "Language and the Development of Spatial Reasoning" from *The Innate Mind* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Spelke, E. (2003) "What Makes Us Smart? Core Knowledge and Natural Language" from *Language in Mind* (eds. D. Gentner, S. Goldin-Meadow). Cambridge: MIT Press, pp. 277-311

Sperber, D. (2005) "Modularity and Relevance" from *The Innate Mind* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Sperber, D., Wilson, D. (1995) *Relevance: Communication and Cognition, 2nd Edition.* Oxford: Blackwell

Sperber, D., Wilson, D. (1996). "Fodor's Frame Problem and Relevance Theory". *Behavioral and Brain Sciences* 19(3): pp. 530-532

Todd, P., Gigerenzer, G. (1999) "What We Have Learned (So Far)" from *Simple Heuristics that Make Us Smart* (eds. Gigerenzer & Todd). New York: Oxford Univ. Press, pp. 357-366

Tooby, J., Cosmides, L., Barrett, C. (2005) "Resolving the Debate on Innate Ideas" from *The Innate Mind* (ed. Carruthers, Lawrence, Stich). Oxford: Oxford University Press

Wason, P.C. (1968) "Reasoning about a rule" *Quarterly Journal of Experimental Psychology 20,* pp. 273-281

Wilson, D., and Sperber, D. (1981) "On Grice's theory of conversation" from *Conversation and Discourse,* (ed. P. Werth). New York: St. Martins Press, pp. 155–78

Wegner, D. M. (2002) *The Illusion of Conscious Will.* Cambridge: MIT Press

Wittgenstein, L. (1953) *Philosophical Investigations (eds.* G.E.M. Anscombe, R. Rhees *trans.* G.E.M. Anscombe). Oxford: Blackwell *(cited as PI)*

Wittgenstein, L. (1967) *Zettel.* ed. G.E.M. Anscombe, G.H. von Wright, trans. G.E.M. Anscombe, Basil Blackwell: Oxford

Wittgenstein, L. (1976) *Lectures on the Foundations of Mathematics.* Cornell University Press: Ithaca,