Framing the web:
Cognitive modularity and the limits of belief revision

Robert Stephens

Department of Philosophy
McGill University, Montreal

August 2014

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

Belief revision practices *ought* to respect the principles of coherence, according to standard norms of rationality. Yet numerous empirical studies suggest our belief revision practices fall hopelessly short of this goal. Worse, a number of influential accounts in cognitive science note that there are hard computational limits involved in any sort of holistic, global belief revision. We are faced with what cognitive scientists call the *frame problem*, which alludes to the difficult question of where to stop considering evidence before committing to (or rejecting) any given belief, yet at the same time, trapped in what Cherniak (1986) refers to as the *finitary predicament* of having limited time and computational resources to engage in the process.

I argue that an effective way to escape this dilemma is to invoke a modular cognitive architecture, where belief revision practices are sub-served, mediated, and heavily circumscribed by informationally encapsulated cognitive mechanisms and heuristic processing. A number of influential accounts have emerged in recent years arguing for such "massively modular" systems as a response to various aspects of the *frame problem* (Carruthers, 2006a; Jackendoff, 2007; Sperber, 2005; Barrett & Kurzban, 2006). I defend my own version of such an account, with a specific emphasis on the question of belief revision within such a modular framework. I begin by exploring the Fodor's (1983) thesis of perceptual modularity and then elaborate the idea, arguing that there is evidence for *assembled* modular structures, including integrative modular assemblies that can execute belief revision processes in a computationally tractable fashion, despite Fodor's well-known objections to this extension of his theory.

I describe a modular, heuristically driven cognitive system that is plausibly capable of *approximating* the sort of global, holistic belief revision practices that rationality demands, while maintaining computational tractability. The price of such a system, however, is that it is *error-prone*—it will have systematic patterns of breakdown, where some beliefs will turn to out to be essentially unrevisable and some inconsistencies of belief will be irremediable. I argue that this prediction of the account is confirmed by current research on memory distortion and delusion. Finally, I demonstrate how my account may illuminate and help resolve some ongoing debates regarding the etiology, doxastic status, and potential treatment of certain monothematic delusions.

Résumé

Les pratiques de révision des croyances doivent respecter les principes de cohérence, selon les normes de la rationalité. Pourtant, de nombreuses études empiriques suggèrent que nos pratiques de révision des croyances ne respectent clairement pas cet objectif. Pire encore, un certain nombre de théories influentes dans les sciences cognitives notent qu'il y a des limites de calcul formidables impliqués dans toute sorte de révision holistique de croyance. Nous sommes confrontés au *problème de cadre*, qui renvoie à la question difficile de savoir où arrêter l'examen des preuves avant de s'engager à (ou de rejeter) une croyance, mais en même temps, pris au piège dans ce que Cherniak (1986) appelle *la situation finie* d'ayant peu de temps et de ressources de calcul à dédier à ce processus.

Je soutiens qu'un moyen efficace d'échapper à ce dilemme est d'invoquer une architecture cognitive modulaire, où les pratiques de révision des croyances sont sous-desservies et fortement encadrées par des mécanismes cognitifs encapsulés ainsi que le traitement heuristique. De nombreuses théories influentes qui ont émergé au cours des dernières années défendent les systèmes "massivement modulaires" comme réponse à divers aspects du problème de cadre (Carruthers, 2006a; Jackendoff, 2007; Sperber, 2005; Barrett & Kurzban, 2006). Je défends ma propre version d'une telle théorie, avec un accent particulier sur la question de la révision des croyances dans un cadre modulaire. Je commence par explorer la thèse de la modularité de Jerry Fodor (1983), puis j'élabore l'idée, soutenant qu'il existe des preuves de structures modulaires assemblées, y compris les assemblages modulaires intégrées qui peuvent exécuter des processus de révision des croyances dans un mode de calcul tractable, malgré les objections de Fodor à cette extension de sa théorie.

Je décris un système cognitif modulaire, entraîné par des processus heuristiques, qui est probablement capable de rapprocher les pratiques de révision de croyance holistique exigées par la rationalité, tout en conservant tractabilité informatique. Toutefois, le prix d'un système de croyance est qu'il est une source d'erreurs et qu'il y aura des mouvements systématiques de rupture, où certaines croyances sont essentiellement non révisables et certaines incohérences seront irrémédiables. Je soutiens que cette prédiction de la thèse est confirmée par de nombreuses études empiriques sur les distorsions de la mémoire et croyances délirantes. Finalement, je démontre comment ma thèse peut éclairer et aider à résoudre certains débats en cours sur l'étiologie, l'état doxastique et le traitement potentiel de certains délires monothématiques.

Acknowledgements

Many thanks are owed to many people who helped me to see this project through to fruition. Foremost thanks go to my supervisor, Ian Gold, whose positivity, patience, and insightful critical commentary made this thesis possible. Much gratitude is similarly owed to Jim McGilvray, whose incisive comments on earlier drafts shaped many of the ideas in this dissertation for the better. I have also had the benefit of stimulating conversations and courses with many members of the Philosophy Department, as well as my fellow graduate students. A large debt is also owed to the Department support staff, Mylissa, Angela, Claudine, and Saleema, who have been immeasurably helpful navigating through the program and the University. Finally, I gratefully acknowledge the assistance received from the Department of Philosophy at McGill, and the Social Sciences and Humanities Research Council of Canada.

Personally I owe a special debt of gratitude to my family for their support, and to Ilana especially for letting me try half-baked ideas out on her, and inspiring me to go back to graduate school in the first place. Finally, I should note that this project might never have been finished it were not for my many friends and neighbours who never failed to ask "how is your thesis coming along?" thus propelling me ever forward to completion.

Table of contents

Abstract ii

Résumé iii

Acknowledgements iv

List of figures viii

Introduction 1

PART I

1 Belief revision: what *is* vs. what *ought* to be 7
1.1 Belief and rationality 8
1.2 Principles of belief revision 10
1.2.1 Coherence and consistency 10
1.2.2 *Change in View* 15
1.3 Dual process views 19
1.3.1 Belief/acceptance 20
1.3.2 Belief/alief 21
1.3.3 The 'Spinozan' and the 'Cartesian' 24
1.3.4 Dual process theories 27
1.4 Review and look ahead 30
Notes for chapter 1 31

2 The 'finitary predicament' 34
2.1 Minimal rationality 34
2.2 Bounded rationality 38
2.3 The frame problem 39
2.3.1 To be(lieve) or not to be(lieve) 40
2.3.2 The Fodorian iteration of the frame problem 43
2.4 Review and look ahead 46
Notes for chapter 2 46

PART II

3 Local modularity 49
3.1 Modularity at the sensory periphery 51
3.1.1 Fodorian modules 53
3.1.2 Why modularity? 57
3.1.3 Empirical evidence of modularity 59

3.2 Arguments and replies 68
3.2.1 Prinz's critique of modularity in general 69
3.2.2 Cognitive penetrability 71
3.2.3 Explaining (away) cognitive penetration 75
3.3 Integrative modularity 78
3.4 Barsalou's simulation theory 85
3.5 Review and look ahead 87
Notes for chapter 3 89

4 Framing beyond the periphery 93
4.1 Assembled modularity 95
4.1.1 If it *looks* like a duck, *walks* like a duck, *quacks* like a duck... 97
4.1.2 A theory of mind module 100
4.1.3 A cheater detection module 102
4.2 Fodor's challenge 104
4.3 Review and look ahead 110
Notes for chapter 4 111

5 Concepts, Belief, and Fodor *vs.* Fodor 114
5.1 Modular concept acquisition 115
5.1.1 Fodor's LOT and conceptual atomism 116
5.1.2 Attractor landscapes—whirlpools of the mind? 120
5.1.3 Modular 'whirlpools'? 125
5.2 Modular concept organization 129
5.2.1 Self-organizing concepts 130
5.2.2 Barsalou's conceptual frames 132
5.2.3 Memory encoding and retrieval as modular processes 134
5.3 Modular belief 139
5.3.1 Associative processing: how far can it take us? 139
5.3.2 Belief revision and the limits of recall 143
5.3.3 Systematic patterns of breakdown 146
5.4 Review and look ahead 147
Notes for chapter 5 149

6 Globality on a budget 152
6.1 The global workspace 153
6.1.1 Blackboard architecture 153
6.1.2 Semantic promiscuity 158
6.1.3 Taking stock of the global workspace 160
6.1.4 Watson 163

6.2 Heuristic approximations 167
6.2.1 Heuristics & biases 168
6.2.2 Test case, or: how I learned to stop worrying and love my new TV 174
6.2.3 Is there a frame problem regarding heuristic selection? 176
6.3 Dumb as a bag of hammers 180
6.4 Review and look ahead 183
Notes for chapter 6 186

PART III

7 Memory distortion 189
7.1 Manipulating memory 191
7.1.1 The misinformation effect 191
7.1.2 Rich false memory 195
7.1.3 Discrepancy detection 197
7.2 Intentional forgetting 199
7.2.1 The DF experimental paradigm 200
7.2.2 When *can*—and when *can't*—we forget? 203
7.3 The bell that can't be unrung 206
7.4 Review and look ahead 208
Notes for chapter 7 209

8 Delusion 212
8.1 Theoretical disputes 213
8.1.1 Doxastic or no? 213
8.1.2 Explanation or endorsement? 217
8.1.3 One factor or two? 219
8.2 Tale of two delusions 222
8.2.1 Modularity and monothematic delusion 223
8.2.2 The Capgras delusion 225
8.2.3 The mirrored-self misidentification delusion 232
8.3 Limitations of a modular account of delusion 238
8.4 Wrapping up 240
Notes for chapter 8 240

Conclusion 245

Works cited 251

List of figures

Fig. 1: The checker shadow illusion. 50
Fig. 2: The Müller-Lyer illusion. 55
Fig. 3: A real world Müller-Lyer. 56
Fig. 4: A Kanizsa Triangle. 59
Fig. 5: The longest arm. 60
Fig. 6: The letter crowding effect. 63
Fig. 7: A speech perception integration "module" 81
Fig. 8: Fodor's "input problem". 105
Fig. 9: Fodor's concept locking process. 123
Fig. 10: The person recognition integration "module". 228
Fig. 11: The mirrored self recognition "module". 235

Introduction

Most philosophical discussions of human rationality and belief revision focus on the normative dimension—the question of how, and why, and under what conditions, we *ought* to believe (or disbelieve) certain things. Psychologists, on the other hand, generally focus their inquiries on the descriptive dimension—the question of how, and why, and under what conditions we *actually do* believe (or disbelieve). There is much crossover, of course, especially in the cognitive sciences, and that is the terrain that this dissertation will explore. My goal is somewhat deceptively simple: I want to trace out an account of the cognitive architecture that human rationality, belief, and belief revision requires—not what it *normatively* requires, insofar as how cognition would have to be structured in order to meet the normative demands of rationality, but, rather, what actual belief revision *descriptively* requires, or appears to require. Part of this job will be diagnostic, i.e., figuring out what we actually *do* in the process of belief revision, and whether it comes anywhere close to what philosophers, for example, tell us we *ought* to be doing. Another part of this job will be reverse-engineering, i.e., working backwards from what we know about how we *do* manage belief to an account of the underlying structure that might make this possible, and what limitations must be respected and may need to be imposed as a result. A final part of the job will be to test the proposed architecture by seeing what predictions it might make regarding belief revision practices, and checking to see if these are confirmed by empirical evidence.

My overarching thesis can be stated simply, though making the case for it will be admittedly complicated, and will move in steps that may seem disjointed at times, rather than as a step-wise logical argument. The thesis is this: belief revision practices must be sub-served, mediated, and heavily circumscribed by modular cognitive mechanisms and heuristic processing. The payoff of a system of belief revision that is mediated entirely by subdoxastic modular functioning is that it is *tractable*—it can actually *work* and get its work *done*—whereas, a system that operates according to the way philosophers tend to talk about belief revision *cannot* work. The price of a system of belief revision that is mediated entirely by subdoxastic modular functioning is that it is *error-prone*— it will have systematic patterns of

breakdown, and will be technically incapable of meeting the globally coherent, holistic principles of belief maintenance typically demanded by norms of rationality. The upshot of this is that we probably should accept a much more deflationary understanding of those norms and principles. In the final chapter of this dissertation I will attack that question head-on, as I will argue that certain cases of monothematic delusional belief—which are paradigm examples of *irrationality*—tend to be irremediable precisely because they are the product of a massively parallel, heuristically-driven, modular system working as designed. Inconsistency, self-deception, cognitive or implicit bias, and even delusion are all just the inevitable result of a system that is optimized for tractability over precision. I will lay the argument out as follows:

PART I *maps out the terrain, the prescriptions and the problems I will tackle:*

In chapter 1, I map out the terrain briefly with regard to how philosophers tend to account for belief revision. I begin by examining Quine & Ullian's *The Web of Belief* as prototypical of a standard *normative* account of belief revision that highlights coherence and conservatism as the fundamental virtues of belief management. I will note that this idealization is in fairly direct conflict with what we actually seem to *do* when it comes to managing our "web of belief"—and that many of the prescriptions inherent in this sort of "Quinean" holism are likely *impossible* to actually adhere to. In this chapter, I also look at other accounts that acknowledge and incorporate these practical limits into an account of belief revision, to varying degrees. I will go through a number of commonly supported and insightful views which move farther and farther from the more idealistic account to propose systems with some sort of limitations, or deflationary expectations, or new doxastic categories and distinctions to try and isolate the problem cases of inconsistency, incoherence, and perseverant false belief. I will argue that all of the accounts in this chapter—all of which are representative of well-thought-out and useful strategies for bridging the normative-descriptive gap regarding belief revision—ultimately fall short of respecting the formidable limits of cognition as mediated by *physical, limited* systems.

Chapter 2 starts from the question of limits, citing Cherniak (1986) on the 'finitary predicament' of human cognition—the limited time and resources we have to devote to the seemingly insurmountable computational tasks that holistic belief revision and inferential

thought demand. I will expand on his concerns by introducing what is known in computing and cognitive science as the *frame problem:* the question of how a system can *frame* a potentially infinite task ahead of time in order to make it tractable. The fact is, we humans already appear to have "solved" this problem: we *do it*—we engage in belief revision, inferential reasoning, even novel creative thought. We might not do it *perfectly*, but we do it. And given the computational challenge of doing it, the human mind must have some pretty good "framing" tricks in order to impose tractability. The effort to create *machines* capable of what human minds are capable of has proven difficult, largely because of this frame problem. Trying to reverse-engineer *how* we get around it will tell us a great deal about how the mind is structured. I will argue that the best prospect we have for a system that can tractably achieve what we can achieve is a *modular* one. Modularity can both explain how we *succeed*, despite heavy computational odds, and it can also explain the many ways in which we *fail*—the ways in which our belief revision practices and reasoning abilities leave something to be desired, and are prone to systematic patterns of breakdown.

PART II *is where I make a positive argument for certain proposals.*

Chapters 3 and 4 make the case for modularity in a fairly step-wise fashion. In chapter 3, I will begin with the fairly *un*controversial thesis that modularity of processing at the sensory periphery is the best way to explain how our sensory perceptual apparati are capable of solving the ill-posed, inverse problem they are faced with: representing an infinitely detailed, constantly shifting multidimensional world in a clear, relatively stable (syntactically encoded) form that the brain can *use.* I will present the modularity thesis based on the work of Fodor (1983), Marr (1982), and Pylyshyn (1984) which posits that sensory perception is mediated by hardwired, automatic, informationally encapsulated, domain-specific, "black box" processing devices, that are cognitively impenetrable to higher-level, top-down cognitive influences. I will relate some empirical evidence supporting the modularity thesis, and look at some objections to it: namely, the objection that there are numerous cases of so-called "cognitive penetration" in which the informational encapsulation of the purported modules appears to be routinely violated. I will defend modularity against these objections. I will present an original argument that the evidence of cognitive

penetration often used against sensory modularity, rather than serving as proof against it, actually just proves that modularity extends *beyond* the periphery to include modular sensory integration functions. In the remainder of chapter 3, I defend that proposal, highlighting numerous cross-modal sensory "illusions" that I will argue are best explained if we presume the existence of integrative modules.

In chapter 4, I expand on the integrative modularity thesis to consider the idea of "assembled" modules that perform higher-level cognitive functions, beyond the sensory periphery, taking the representational output of perceptual modules as their proprietary input. This is a more controversial thesis, and Fodor himself, who is largely responsible for the initial modularity thesis, thinks "assembled" modules are a non-starter. I will introduce a few plausible-seeming candidates of assembled modules in the realm of social cognition, and use them as test cases against Fodor's objections. I will show that Fodor's objections can be met, and that assembled modularity not only fits with the original modularity thesis comfortably, but that the empirical evidence supports it. Furthermore, I will note some so-called *massively modular* expansions of the modularity thesis, and engage Fodor's objection that a more massive construal of modularity, specifically modularity of belief revision and inference practice, must fail because these practices demand "Quinean", isotropic processes that modules are incapable of.

In chapter 5, I will expand on that argument in order to explain why Fodor's concern is misplaced, and that his own theory of concept acquisition can be used to explain why he is wrong on this point. I endeavor to show that Fodor's (2008) account of concept acquisition is much more explanatory if it assumes assembled and integrative modular structures to subserve the concept locking process that he describes. Additionally, I note how on Fodor's account of concepts as mental "files", concept acquisition should result in a self-organized compartmentalized storage arrangement which allows for fast, associative searching later on. In short, Fodor's account of concepts contains key elements to resolving the frame problem, despite the fact that he doesn't view it this way himself. Also in chapter 5, I introduce the pivotal research of psychologist Endel Tulving on memory retrieval to explain how it dovetails nicely with both Fodor's account of concepts, and the account I defend regarding how quick, effective, associative retrieval processes, subserved by modular acquisition and filing sub-routines, can render belief revision and relevance determination tractable.

Chapter 6 is where I will present the bulk of my own positive account of how assembled and integrative modularity can work to underwrite a system of belief revision and rational deliberation in general. I will need to look at 2 elements in order to sketch this account:

1. (Chapter 6.1) What sort of *global workspace* is available for bringing disparate systems into contact with one another.

2. (Chapter 6.2) What sort of heuristic algorithms can expedite and/or limit searches and judgment procedures.

My goal here will be to bring together insights from a number of different thinkers, as well as evidence from numerous studies on cognitive bias, to sketch a picture of how cognition might be structured in a way that *approximates* the norms of belief revision and rational thought, while still maintaining plausible computational tractability and skirting the *frame problem*. The latter half of chapter 6 brings on board the many theoretical resources uncovered and established by the "heuristic & biases" research program, beginning with Tversky & Kahneman (1973). I will show how the invocation of heuristic search and judgment processes, combined with what I discuss in chapters 5 & 6, suggests a variation on a massively modular account of belief revision and inferential practice—one that is grounded in empirical research, and requires a great deal less "hand-waving" as to how it works, in comparison to other, similar proposals put forth in recent years. My ultimate conclusion in PART II is in line with the "bounded rationality" thesis (Gigerenzer & Todd, 1999) that the human mind is equipped with an "adaptive toolkit" of heuristic algorithms that approximate holistic processing. However, I will offer a slight revision to the metaphor: the "toolkit" is really just a *bag of hammers*—brute, not ideal for every job, but generally effective at most. However, the tractability bonus of a hammer is priceless: because to a person with a hammer, everything becomes a nail. My argument is that our minds are designed to *transform the problems* we are faced with into solvable formats. So we are, in the end, as dumb (or as smart) as a bag of hammers.

PART III *is where I turn to empirical findings for consolidation, confirmation, and support:*

In chapter 7, I look to empirical research regarding *memory distortion* and *belief perseverance* to test whether the proposals I have made in PART II are supported by the

psychological evidence. I will argue that this evidence *does* clearly support the sort of massively parallel, associative, heuristic-based, modularly mediated, integrative processing account I have presented. I will also show how research into false memory, implicit memory, and the conditions under which one can (and more importantly *cannot*) intentionally *forget* things, all support the fundamentally modular—viz., automated, domain-specific, and informationally encapsulated—architecture I have sketched out in previous chapters.

Chapter 8 serves as further empirical testing of predictions that fall out of my account. In this final chapter I will look specifically at monothematic delusional belief. I will argue that these sorts of delusions are precisely the *systematic patterns of breakdown* that one should expect from a reasoning and belief revision system underwritten by a massively modular architecture. In short, delusions are to the reasoning system what cross-modal illusions are to the perceptual system—I will argue that the analogy is direct and dispositive. I will look specifically at two delusional syndromes—the Capgras delusion and the mirrored-self misidentification delusion—to show that not only their etiology, but also their apparent incorrigibility, can be explained using my modular account of belief revision. I will also suggest that this account can help settle some ongoing questions in the literature on delusion in general, and help resolve some particular puzzles about why delusional beliefs fail in many ways to act like belief more generally. I will show how the account of modular belief revision I have sketched provides a fairly elegant solution to some of those puzzles.

PART I

1 Belief revision—what is *vs.* what ought to be

"There is no simple touchstone for responsible belief"
(Quine & Ullian, 1978:8).

My target in this dissertation is believ*ing* (and even more, *un*believing). I am not undertaking an argument about the nature and/or status of belief, epistemological, metaphysical, or otherwise. Rather, I am going to be making claims about the cognitive systems that underwrite belief revision. I will be tackling a question about cognitive architecture: given what we know *empirically* about our processes of reasoning and belief revision—both in our successes and, more relevantly, our systematic failures—what sort of system could, in principle, account for that evidence?

The questions just listed, of course, all point toward the construction of a *descriptive* account of belief revision. Of course, any account of belief revision is also going to have a certain *normative* dimension as well. How one goes about revising one's beliefs will be *judged* by others—it's a job that we think can be done well or poorly. When a person does an *especially* poor job of revising his or her beliefs—clinging to clearly false beliefs, not considering evidence appropriately, believing contradictory propositions, etc.—we consider it evidence that the person in question is *irrational*. We expect them to *do better.* And we expect not just that they revise beliefs "better", we expect that their behaviour, and attitudes, and preferences will reflect those revisions: we won't consider them "rational" if their actions or attitudes belie their professed beliefs, or if they routinely violate transitivity in preference or evaluation, or if their revision strategies seem too arbitrary (i.e., it won't be enough to believe the right thing; it must be believed for the right reasons).

In this chapter, before we get to the descriptive account of belief revision that will comprise the bulk of this dissertation, I want to quickly survey the philosophical landscape

regarding belief revision—starting from the normative side of the question. I have chosen to begin with Quine & Ullian's *Web of Belief*, specifically because it suggests an idealized view of belief revision, one that fits intuitively with what we probably *all* think belief revision practice *ought* to be: an ongoing project of coherence and consistency-checking. Quine & Ullian's account isn't unique in this sense, but I have selected to focus on it for another reason: Quine will serve as a stalking horse throughout the first 6 chapters of this thesis. Indeed, as we will see below, a lot of the literature on belief revision in cognitive science refers to the "Quinean"[1] nature of belief, insofar as belief appears to be functionally *isotropic*, which is to say that belief revision demands holistic global access to the epistemic background of one's cognitive system. Jerry Fodor—a self-professed Quinean—explains the isotropy of belief as follows: "the level of acceptance of any belief is sensitive to the level of acceptance of any other and to global properties of the field of beliefs taken collectively" (Fodor 1983: 110). After Quine's account, I will look at a few further accounts, which stray farther and farther from the idealized prescriptive account of belief revision, and which attempt to account for our evident and numerous failures to revise belief in anything like the way a normative account demands we ought to. Before we get to Quine & Ullian, however, a few words on the link between belief and rationality.

1.1 Belief and rationality

Managing one's beliefs is a pretty heavy responsibility, as one's claim to rationality generally hinges on doing it right. John Broome, is his discussion of the normativity of rationality takes the following as a starting position:

> Rationality requires various things of you. For example, it requires you not to have contradictory beliefs, to believe what follows by modus ponens from things you believe, to intend what you believe to be a necessary means to an end that you intend, and to intend to do what you believe you ought to do. (Broome, 2007: 161)

The relationship between belief and rationality is intertwined, insofar as each tends to be the yardstick via which the other is measured. The "requirements" of rationality mentioned by Broome in the quote above all highlight various expectations we project onto the holder of belief, which speak to some clear divisions we can immediately suggest as to various dimensions of rationality. I will follow the divisions laid out by Lisa Bortolotti (2010), here, for two reasons: 1) because I think her tripartite analysis is a simple and clear account of the

ways we expect belief and rationality to coexist; and 2) because the *last* chapter of this dissertation will focus on the question of *delusional* beliefs—irrational belief *par excellence*—and Bortolotti's account will loom largely over that chapter.

Bortolotti suggests there are three dimensions of belief that relate directly to rational norms: *procedural, epistemic,* and *agential*:

> 1) Beliefs have relations with the subject's others beliefs and other intentional states.
>
> 2) Beliefs are sensitive to the evidence available to the subject.
>
> 3) Beliefs are manifested in the subject's behaviour. (Bortolotti, 2010: 12)

As a result, she proposes that we should view rationality as having three associated dimensions. *Procedural rationality* refers to the appropriate integration of beliefs within a coherent system; *epistemic rationality* demands that beliefs are properly supported by evidence, and are responsive to evidence; finally, *agential rationality* has to do with an agent's ability to think and act in a way consistent with her beliefs, and her actions and thought should be explicable in terms of those beliefs (Bortolotti, 2010: 14). Bortolotti's divisions are not meant to be taken as exhaustive, or even as exemplary of standard accounts of rationality. Indeed, on many accounts, all three of Bortolotti's dimensions of rationality would fall, at least in part, under the rubric of *instrumental rationality*. But I am not giving an analysis of rationality here. The primary focus of this dissertation is belief revision—and as we shall see, the connection to "rationality", on any dimension, will turn out to be a fair bit more complicated and ambiguous than the initial discussion presumes it to be.

For instance, the "demands" of procedural, epistemic, and agential rationality may be stricter than we are capable of meeting. The reality is that much of our belief-based behaviour may not coincide with what's ostensibly "rational". But then, when belief falls outside of these rationality constraints, what do we want to say about that? Do we want to deny the ascription of "belief" status to those beliefs that fail to abide by the norms of rationality? Or do we accept that we are irrational? (Or is it that we *are* rational, but have merely *acted* irrationally?) Maybe we simply take the norms of rationality as a *regulative ideal*, and assume that, following Socrates, if we *knew* better, we'd *do* better.[2]

Regardless of how we want to answer these questions, we are going to run into problems with our definitions and expectations on both sides of the belief-rationality relation, as we shall see throughout this dissertation. Some points will put pressure on the definition

of *belief*, as in many instances what we intuitively want to class as "belief" may fail to fulfill the role(s) assigned to it by our understanding of rationality. At the same time, many cases of unambiguous *belief* will in turn clash with our intuitive notions of rationality. In such cases, if something has to give, I will side with Bortolotti's analysis: that rationality should not be a constraint on the ascription of belief—i.e., we should adjust our notion of rationality to fit our notion of belief, not the converse. By the end of this chapter, we will get a better sense for how best to understand the role of and relation between belief and rationality. First, let's begin with the ideal picture: how *ought* we manage and revise our beliefs in a way that will also us to remain within the domain of the *rational?*

1.2 Principles of belief revision

In this section I will review two influential philosophical accounts of belief revision—Quine & Ullian's (1978) *Web of Belief*; and Harman's (1986) *Change in View*—in an attempt to highlight the normative dimension most philosophical discussion of rationality and belief revision practice reveals, as well as some descriptive realities and constraints on actual day-to-day belief revision strategies that point to the central problem addressed in this dissertation. By no means is the discussion in this chapter meant to be exhaustive with regards to accounts of belief revision—my goal is merely to trace out some of the terrain, and reveal some of the tension between how philosophers tend to suggest belief revision *ought* to go, and how it *actually* goes in many cases. Along the way, I will highlight some lurking problems and puzzles, and issue a few promissory notes regarding sections to come later in this dissertation where those problems and puzzles will be taken up again.

1.2.1 – Coherence and consistency

In *The Web of Belief*, Quine & Ullian lay out an account of "many of the principles by which reasonable belief may be discriminated from unreasonable belief" (1978:8). Their approach to rationality and belief revision espouses principles similar to those that are invoked as criteria of adequacy in theory evaluation in science. On this view, individual rationality is of a piece with scientific theory construction: both aim at a veridical representation of reality, and the use of inferential procedures to aid in the recognition of patterns that allow for successful prediction based on past observation. As Quine notes in 'Two Dogmas of

Empiricism': "the conceptual scheme of science [is] a tool, ultimately, for predicting future experience in the light of past experience" (1951:41). For Quine & Ullian, belief formation is essentially hypothesis formation, and hence should abide by similar constraints and according to similar virtues:

> Calling a belief a hypothesis says nothing as to what the belief is about, how firmly it is held, or how well founded it is. Calling it a hypothesis suggests rather what sort of reason we have for adopting or entertaining it… Hypothesis, where successful, is a two-way street, extending back to explain the past and forward to predict the future. What we try to do in framing hypotheses is to explain some otherwise unexplained happenings by inventing a plausible story, a plausible description or history of relevant portions of the world. What counts in favor of a hypothesis is a question not to be lightly answered. We may note five virtues that a hypothesis may enjoy in varying degrees. (1978:66)

The five virtues in question are *conservatism, modesty, simplicity, generality,* and *refutability*—which will sound familiar to philosophers of science (*cf.* Popper, 1953/1988).

The fact that *conservatism* is the *first* virtue on Quine & Ullian's list should be a clear indication of the account of rational belief revision that they are giving—an account in which coherence and consistency-checking will be the primary tool in the belief revision kit. The "web" of belief is structurally unsound if contradictory beliefs are held. Indeed, on pain of irrationality:

> we can no longer believe all of a set of sentences to be true once we know them to be in contradiction with one another ... Once we recognize a conflict among our beliefs, it is up to us to gather and assess our evidence with a view to weeding out one or another of the conflicting beliefs. (1978: 14)

Note the stress on "recognition" in the passage just quoted: one cannot root out inconsistency without first identifying the inconsistency as such. This is the hard part of belief revision, as "inconsistency is not always obvious" (11) and "the reason why widespread misbelief can thrive is that the ignorance of relevant truths is often accompanied by ignorance of their ignorance" (59).

So how do we remedy the ignorance of ignorance? It's not fully clear what Quine & Ullian think about that, actually, as their discussion moves quickly back to pointing out that one should not "believe the impossible" which is precisely what believing a contradiction would entail:

> One can't believe a thing if one sees that it is impossible.... When conflicts arise, creating impossible combinations, we cannot rest with them; we have to resolve them. (60)

But, again, this requires first recognizing the conflict. Note that this particular problem—we will provisionally call it the problem of *inconsistency awareness*— is one that will be readdressed repeatedly and at length later in this dissertation, as it poses a problem for standard accounts of belief revision. In short, there is reason to suspect that *inconsistency awareness* may be beyond the capacity of human cognition in a large number (perhaps a majority) of cases in which inconsistency of belief is at play—even when there is unambiguous evidence of the inconsistency. What that says about the applicability or scope of an account of belief revision that relies on recognizing contradictions within one's belief web as a precondition to remedying them is a concern that we will have to set aside for the moment. Let us, for the time being, grant Quine & Ullian that *in many cases*, we will, indeed, be aware of inconsistent belief sets, so that we may continue looking at the revision procedures they prescribe.

Given the connection established between belief and scientific hypothesis formation, it should be no surprise that Quine & Ullian stress the importance of evidential backing for belief, based primarily on observation, and what can be derived deductively and inductively from direct observation and observed regularities of past evidence. Evidence is what separates belief from mere opinion, as "[a] person need never have assessed the evidence for anything in order to be rich in opinion" (14), and therefore evidence-sifting is crucial to belief formation, though Quine & Ullian are careful to distinguish the *evidence for* a particular belief from the *cause* of a belief. Often the two coincide, but not always, and one must be on the lookout for beliefs that have been *caused* without proper evidential backing. One supposes this will require a fair degree of vigilance. Quine & Ullian offer that "[o]ne obvious test of evidence is this: would it still be taken to support the belief if we stripped away all motives for wanting the belief to be true?" (15).[3]

So what counts as good evidence for a belief? Quine and Ullian's view is that "the ultimate evidence that our whole system of beliefs has to answer to consists strictly of our own direct observations—including our observations of our notes and of other people's reports" (21).[4] Of course, in many instances, our observations will be incomplete or mistaken, or we may fail to draw the proper inferences from them, or fail to properly adduce patterns and regularities among observations (either by not noticing the patterns, or mistaking simple coincidence and loose correlation for a meaningful pattern). Quine & Ullian

recognize this—that "naturally we leave many points unchecked" (21)—though they argue that mistakes and contradictions will largely reveal themselves through prediction failures. Highlighting Kuhn (1962), they note that individuals, like scientists, rely on "failures of existing rules [as] the prelude to the search for new ones" (31)—when confuting evidence arises, theory must adjust, though the process of adjustment needs the virtue of *conservatism* at the forefront.

> When a set of beliefs is accumulated to the point of contradiction, find the smallest selection of them you can that still involves contradiction… We can be sure that we are going to have to drop some of the beliefs in that subset, whatever else we do. In reviewing and comparing the evidence for the beliefs in the subset, then, we will find ourselves led down in a rather systematic way to other beliefs in the set. Eventually we will find ourselves dropping some of them too.
>
> In probing the evidence where do we stop? …In practice the probing stops when we are satisfied how best to restore consistency: which ones to discard among the beliefs we have canvassed.
>
> Our adjustment of an inconsistent set of beliefs may be either decisive or indecisive. If it is decisive, each belief of the set is either kept or switched to disbelief. If it is indecisive, some of the beliefs simply give way to non-belief; judgment on them is suspended. (Quine & Ullian, 1978:18)

Here, again, the account requires a fairly maximal level of conscious access to one's beliefs—it demands *inconsistency awareness*, both in the detection of contradiction, and in the amelioration of the set. Quine & Ullian also touch on what will be elaborated below as *the frame problem*—the question of "in probing the evidence, where do we stop?" Quine & Ullian's answer is that we stop when we have best restored consistency. But consider, for a moment, someone with merely 150 'beliefs'—150 factual propositions in his or her epistemic database. Now imagine that person is presented with a novel proposition and has to decide whether to believe or disbelieve it, and incorporate or reject the belief from the database accordingly. If we adhere to a Quinean account of belief revision, then simple *coherence* and *conservatism* demand that the belief needs to be checked against the current epistemic background—in this case the 150 beliefs already in the database. But imagine creating a truth-table to test this set for truth-functional consistency: a set 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, $1/100^{th}$ of a second (which seems unreasonably fast), to check the set for truth-functional consistency would take roughly 9×10^{35} years to complete (just under a billion billion billion billion years).[5] 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. The tractability issues inherent in this *frame problem* will be the subject of chapter 2, below. For now, we will put it aside to finish laying out Quine & Ullians's account of how we ought to revise belief, regardless of whether it's computationally feasible.

As described, belief revision will follow the model of standard abductive inference, or *inference to the best explanation*.[6] C.S. Pierce (1931) is largely responsible for delineating abductive inference as a specific sub-species of inductive inference in general. As Pierce defines it:

> The form of inference, therefore, is this:
> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true. (1931: §5.189)

A, in this example, is the best explanation for *C.* But *A* is essentially a product of non-demonstrable inference—we hit on *A* by asking ourselves for a hypothesis that could explain *C,* which comes with no prior explanation, as it was a surprise—a bit of reverse engineered reasoning. Lipton (2004) elaborates on what the model of abductive inference should comprise, noting that there are competing notions of the 'best explanation' that are often employed: on the one hand, the 'best' explanation may be construed as the one that is best supported by the evidence—the "likeliest" explanation—and, and on the other hand, the 'best' explanation may be the one that affords greatest understanding (evidence notwithstanding)—the "loveliest" explanation (Lipton, 2004:57). Lipton argues that most philosophers (mistakenly) tend towards the "likeliest explanation" view, and look at abductive inference as an evidence gathering and hypothesis-testing affair almost exclusively. He contends that "choosing likeliness would push Inference to the Best Explanation towards triviality… we want our account to give the *symptoms* of likeliness, the features and argument has that lead us to say that the premises make the conclusion likely. A model of Inference to the *Likeliest* Explanation begs these questions" (60).

> The distinction between likeliness and loveliness is, I hope, reasonably clear. Nevertheless, it is easy to see why some philosophers may have conflated them. After all, if Inference to the Loveliest Explanation is a reasonable account, loveliness and likeliness will tend to go together, and indeed loveliness will be a guide to likeliness. Moreover, given the opacity of our 'inference box', we may be aware only of inferring what seems likeliest even if the mechanism works by assessing loveliness. (Lipton, 2004: 61).

The reference here to "opacity" is interesting,[7] as it doesn't sit well with Quine & Ullian's account of belief management and revision, which requires *inconsistency awareness* and clear conscious access to one's beliefs (all of them, in principle, it seems), the evidence that confirms or confutes those beliefs, and the *manner* in which it does so. Lipton's definition of inference to the best explanation suggests that to some extent, Quine & Ullian's prescriptions for rational belief revision are aimed at the wrong target: checking the evidence and weeding out the beliefs that are unsupported may not be the way we in fact *do* proceed when making such inferences, although it may *seem* to us that this is what we are engaged in.

We will return to the question of abductive inference in later chapters at greater length, and will follow the leads of Lipton's argument much further at that time. For now, suffice it to say that the prescriptive program for belief revision and maintenance of a consistent 'web of belief' may not be as straightforward an affair as has been presented by Quine & Ullian. In the next section, I will examine Gilbert Harman's (1986) account of belief revision— *Change in View*—in order to press further the distinction between how we *ought* to revise belief, versus what psychology teaches us about how we actually *do* (or don't) revise belief.

1.2.2 *Change in View*

Harman (1986) agrees, for the most part, with the focus on consistency and conservatism as the primary virtues of a sensible belief revision process, though his position is noticeably less maximalist than Quine & Ullian's, insofar as Harman recognizes a number of practical and psychological limitations that gum up the process of coherence-checking and conscious assessment of evidential reasons for belief.

> Belief revision is like a game in which one tries to make minimal changes that improve one's position. One loses points for every change and gains points for every increase in coherence. One does not normally try to maximize. One tries to get 'satisfactory' improvement in one's score. One 'satisfices' rather than maximizes.[8] (Harman, 1986: 68)

Whereas Quine & Ullian prescribe whittling one's belief sets down to weed out (all) contradictions, Harman appears to be more willing to make some practical tradeoffs, while still highlighting coherence and conservatism as goals. On Harman's view, whatever principles of belief revision we employ, they can't be simply *logical* principles. Although

logical implications may reveal contradictory beliefs in some instances (which is good, as far as it works), if we actually adopt something like a "*Logical Closure Principle"* in general for belief management, then we will be demanding that a coherent belief system will become "cluttered" by the "many trivial things [that] are implied by one's view which it would be worse than pointless to add to what one believes" (12). Harman suggests:

> there is no clearly significant way in which *logic* is especially relevant to reasoning. On the other hand immediate *implication* and immediate *inconsistency* do seem important for reasoning. (1986: 20)

The focus on "immediacy" here is to keep belief revision from being closed under these logical principles – Harman invokes a principle of *"Clutter Avoidance"*—"a metaprinciple that constrains the actual principles of revision ... One way to do this is to accept a new belief *p* only if one has (or ought to have) an interest in whether *p* is true" (15). "Immediate" implication or inconsistency will be circumscribed by context, and the "*Interest Condition*" will delineate the context(s) in which one should reasonably be expected to invoke *implication* and *inconsistency* principles in one's belief revision schema.[9] One's "interest", according to Harman,

> may be simple, unmotivated curiosity, but it will more often arise in accordance with such principles as the following:
>
> *Interest in the Environment* One has a reason to be interested in objects and events in one's immediate environment. (So one fairly automatically notices 'salient occurrences' that are 'right before one's eyes').
>
> *Interest in Facilitating Practical Reasoning* If one desires *E* and believes *M*'s being true would facilitate or hinder *E,* one has a reason to be interested in whether *M* is true.
>
> *Interest in Facilitating Theoretical Reasoning* If one is interested in whether *P* is true and has reason to believe knowing whether *Q* is true would facilitate knowing whether *P* is true, one has a reason to be interested whether *Q* is true. (1986:55)

Additionally, Harman notes that there is a standing interest in not being (immediately) inconsistent in one's beliefs, and moreover, this immediacy needs to incorporate some awareness of where the relations of implications between beliefs lie—or else when faced with apparent (direct) inconsistencies of beliefs, we might go astray in our revision: we might "abandon one of the explicitly competing beliefs without giving up any of the beliefs which imply it" (56). We will need to weed out the beliefs that lead, via immediate implication, to

inconsistency. This will require a twofold awareness of both direct and *indirect* inconsistency (via implication).

I have already noted above in §1.2.1 that the demand of *inconsistency awareness* is one that will feature prominently is this dissertation as a major obstacle to any account of belief revision, and Harman does recognize that there are problems lurking in this regard. In chapter 4 of *Change in View*, Harman detours into a brief discussion of numerous studies on belief perseverance, and the immunity of certain (false or mis-)beliefs to standard revision procedures—largely because there appear to be unconscious or implicit commitments and/or mechanisms that occlude inconsistencies from doxastic awareness. In Chapter 7 of this dissertation, we will explore the implications of this research in much greater detail (and by that point I will have proposed an account which I believe will be confirmed by and help explain the phenomena revealed in these experiments). For now, we will look briefly to the studies Harman highlights in order to lay down a marker regarding what I will be arguing is one of the primary troubles with standard accounts of belief revision: they require an *inconsistency awareness* and conscious remediation efforts which empirical evidence by social and cognitive psychologists in the field of "intentional forgetting" suggest may be humanly *impossible*, and which, instead, lend support to a particular account of modular cognitive architecture (which I lay out in PART II, below).

One quite robust finding is that of Anderson & Ross (1975),[10] in which subjects continued to make judgments based on misinformation, even though they had been fully debriefed that the information was false. Other beliefs and judgments implied by the false belief persevered regardless, and *further ones* were still made based on the information. In the study, subjects were given false information about their abilities to perform a particular task (in this case, detecting and distinguishing false suicide notes from legitimate ones)—subjects were told either that they were significantly better or worse than average in this regard. Their judgments about their own abilities were formed accordingly. Subsequently, they were debriefed and shown incontrovertible evidence that their "performance" on the task was manipulated, and the information they had been given about their "results" was false, and all beliefs based on it were therefore unwarranted. Subsequent self-reports reflected that subjects *maintained* the belief in their "ability" despite the debriefing. As

Anderson & Ross note in a follow-up study (1980 – to be discussed in more detail in chapter 7, below):

> a theory concerning the relationship between two variables—generated through exposure to a minimal data set—can survive even a complete refutation of the formative evidence on which the theory was initially based. (Anderson & Ross, 1980: 1043)

Harman further references a previous study by Anderson & Bower (1973) in which associative links made from false information believed at one point and then subsequently disbelieved will persevere unless they are *positively* undermined with competing associations that can take their place. Similarly, Nisbett & Ross (1980) established through a number of experiments that the debriefing *can be* successful only when the phenomenon of belief perseverance is made salient to subjects *and* if the false information is positively undermined in such a way that all associations, causal explanations and implications that flowed from the initial misinformation are also positively undermined *and replaced with new explanations and/or implications*. (Harman doesn't discuss them in particular, but we will look in chapter 7, below, at numerous additional studies that highlight these belief perseverance, "misinformation effects" and similar phenomena.[11]) For the time being, the important point to recognize—and which Harman recognizes—is that *inconsistency awareness* is more complicated than a standard, normative account of belief revision seems to expect or demand. It isn't enough, apparently, to even be made aware of immediate inconsistencies in the belief set, and eliminate the misbelieved information: one's belief set will still likely be tainted by contradictory beliefs via implications and associations, and, crucially, *will be prone to continued formation of contradictory beliefs.*[12] Over the course of this dissertation I will argue that this is an inescapable consequence of cognitive architecture, and though it may dim the prospects of certain analyses of belief revision that are generally well-received in philosophy, this is not to say there are not still useful normative principles of belief revision that can take the architectural limitations into account.

Harman's account at least recognizes some of the limitations, which explains his reliance on "satisficing" principles—rather than maximizing ones—and his stipulation that conditions of *Interest* and *Clutter Avoidance* be upheld in any account of revision, rather than simple logical principles of implication and inconsistency avoidance.

> Principles for revising what one fully accepts promote conservatism and coherence. Conservatism is reflected in the principle that current beliefs are justified in the

> absence of any special challenge to them and in the principle of clutter avoidance, which limits the newly inferred conclusions to those in which one has a reason to be interested. Coherence is reflected in one's disposition to avoid inconsistency and a tendency to promote explanatory and implicational connections among one's beliefs. (1986: 116).

However, his account still courts the *inconsistency awareness* problem, and actually introduces another, potentially larger problem: if we are to shape our revision practices around "interests", then we are presuming that we will have some way of determining those interests—of context-framing, and assigning appropriate degrees of *relevance* to various beliefs. As I will discuss in more detail in chapter 2, there are computational restraints that will bear on this question, and will suggest that any account of belief revision that proceeds from an assumption that relevance and interests are (always, or even mostly) introspectible is going to run into what we will refer to as the *frame problem:* the problem of determining context or relevance in the first place. Before delving further into that problem, however, let's first look at a few more accounts of belief revision in order to tease out further problems lurking in many intuitively appealing and well-received accounts.

1.3 Dual process views

One way in which we might try to reconcile our normative impulses regarding global doxastic coherence with the stubborn facts of belief perseverance, inconsistency unawareness, and the lack of time and processing power required to introspect one's belief set, is to posit two levels of processing, (or two types, or systems) of "belief"—one that is amenable to conscious evaluation and coherence-checking, and one that is not. In this section we will look quickly at some accounts that offer promise in this regard. First, I will trace out Jonathan Cohen's (1992) proposal that we need to separate *belief* from *acceptance* in a clear cut way if we hope to understand revision practices, and failures thereof. In 1.3.2-1.3.3, I will be highlighting some other accounts, not technically in the "dual process" camp, but which nevertheless discuss belief and belief revision as involving 2 tracks, in such a way that certain issues already highlighted in §1.2 may be "explained away" as artifacts of separate cognitive procedures, with distinct levels of accessibility and introspectibility. In 1.3.2, I will look at an interesting proposal by Tamar Szabo Gendler (2007; 2008) positing a new category of mental state—*alief*—which she argues can help explain some of the

inconsistency awareness problems under discussion. Additionally, in §1.3.3, I will note in passing an interesting idea from psychologist Dan Gilbert (1991; 2002) suggesting that a lot of the empirical evidence on belief revision, belief perseverance, misinformation effects and false memory militates against what he terms the *Cartesian* view—that we evaluate information before believing it—and supports, rather, the *Spinozan* view—that first we must believe a proposition, and only subsequently evaluate it for truth. Finally, I will turn to explicit "dual process" views, including those of Tversky & Kahneman (1974), Evans & Over (1996), Sloman (1996), Stanovich & West (2000), and especially the account of Keith Frankish in *Mind and Supermind* (2004), which presents a theory of belief revision that incorporates many of Cohen's arguments regarding acceptance and premising policies within an empirically supported dual-system view of cognitive architecture. I will argue that such dual process views push in many promising directions, though still court the 'frame problem' under discussion later in chapter 2.

1.3.1 Belief/Acceptance

Cohen (1992) argues that in order to account for common failures of rational belief revision, we need to understand that not all of what we call "believing" is strictly the same sort of activity. Cohen states that we need a sharp distinction between *believing* and *accepting.* On Cohen's account:

> To ***believe*** that p is to have a disposition to treat p as true and not-p as false.
>
> To ***accept*** that p is to have or adopt a policy of deeming, positing, or postulating that p.
> (Cohen 1992: 4)

There are essentially 3 ways in which *belief* and *acceptance* come apart on this view:[13]

1. Belief is *truth-directed*, whereas acceptance is aimed at explanatory utility or practical success. We can't *believe* something without simultaneously taking it to be *true.* However, we can, and often do, *accept* a premise, while remaining agnostic, or in some cases even doubting, its truth (such as when entertaining an unlikely hypothesis in science, thinking through pros and cons in practical reasoning, or preparing a defense for a criminal client).

2. Non-doxastic acceptance is *under voluntary control*—we can *elect* to entertain (and cease to entertain) various premises in our theoretical and practical reasoning.

Whereas one does not seem to choose one's *beliefs*. One generally *finds oneself* holding certain beliefs, and ridding oneself of a belief is not merely a choice, it requires a process.

3. Acceptance is *context-relative*, insofar as an accepted premise in one context does not necessarily carry over to another (e.g., the lawyer, again, who accepts her client's innocence while on the clock, but believes her client guilty outside of that context, when she is home, eating dinner). Belief, on the other hand, is generally context-independent.

On this view, much of our conscious *belief* revision processes will end up mediated by way of *acceptances*—especially in the realms of inference to the best explanation, and coherence checking in general: we use *premising policies* as precursors to belief, in order to evaluate propositions, evidence and arguments. A premising policy is simply a conscious decision to treat a given proposition as provisionally "true" for the purpose of employing it in reasoning, or "testing it out". Often, there is a point where acceptances can transform into beliefs, after crossing some sort of inferential or evidential threshold (e.g., in scientific hypothesis testing, what begins as an acceptance can graduate to belief). This, Cohen argues, gives us an explanation of how we may end up with inconsistent beliefs sets, or actions not in accord with beliefs: these are simply occasions of "acceptance dominating belief" (Cohen, 1992: 141). By this, Cohen means that it may be quite easy to slip straight from the acceptance to belief without properly advancing through the steps (e.g., without having properly examined the evidence, or the logical strength of the arguments, etc.). Alternatively, under certain circumstances, we may essentially *forget* that a certain proposition was only *accepted*, provisionally, under a premising policy, in order to effect some bit of reasoning; we may then subsequently mistake the (merely) accepted proposition for a full-fledged *belief.*

1.3.2 – Belief and alief

Tamar Szabo Gendler has recently written a number of papers on a proposed category of mental state—*alief:* "an innate or habitual propensity to respond to an apparent stimulus in a particular way" (2008a: 553)—which could she argues could help us make sense of a number of the belief maintenance and revision issues we are exploring in this chapter, specifically the

cases where non-conscious attitudes seem to infect actions that are putatively engaged under contradictory conscious attitudes. She begins with examples such as the "Skywalk" over the Grand Canyon, or the Willis Tower "skydeck" in Chicago, where one can stand hundreds of feet above the ground on a transparent surface: in this case, one surely *believes* that one is safe, and yet one's autonomic nervous system reacts as if one believes on is in peril.[14] What causes the trembling? Do we both believe we are safe *and* believe we are unsafe?[15] If so, that is a straight up contradiction, which we should resolve using evidence. But then, on a repeat visit to the skywalk (or the cage) we should *not* tremble. Yet we do. Gendler suggests the cause of the trembling can't be *belief*, but rather an implicit belief-like state she terms *alief*.

> ***a***lief is: ***a***rational, ***a***utomatic, ***a***ssociative, cognitively ***a***ntecedent to other attitudes, ***a***ction-generating, ***a***ffect-laden, shared with ***a***nimals... Paradigmatic alief is at least a four-place relation, [though] it is tempting to slip into the more natural two-place usage ... The suspended man alieves (all at once): high up above the ground right now, dangerous scary place to be, tremble. (2008a: 557-9)

As further support, Gendler (2008a;2008b) turns to studies done by Paul Rozin and colleagues on "sympathetic magical thinking" (1986; 1992; 2003), in which subjects express clear aversions to eating chocolate shaped like dog feces, or benign water labeled as poison, or food touched by a sterilized cockroach, despite clear and unambiguous debriefing as to the perfect safety for consumption of the items in question. Amazingly, in Rozin & Tuorila (1993), subjects express an aversion to drinking from a bottle merely labeled "*not sodium cyanide".*[16] Rozin & Nemeroff (2002) explain that:

> people are usually either aware, or can easily be made aware, of the 'irrational' aspects of these laws [of magical contagion]. Thus, when educated Americans refuse to eat chocolate shaped into realistic-looking dog feces, or refuse to eat food touched by a sterilized cockroach, they are actually aware that this 'makes no sense', yet acknowledge their feeling of aversion. They can often not overcome this aversion and 'be rational'. (Rozin & Nemeroff, 2002: 202)

Gendler suggests that in these cases, subjects *believe* the food to be edible and germ-free, and the water to be potable, but their *aliefs* lead to an aversive response. There isn't, properly speaking, an inconsistency of *belief* here under this description, and for Gendler, this is an important motivating factor for separating alief from belief, as:

> there is a distinct role that the notion of belief needs to play in our cognitive repertoire if it is to bear the relation to knowledge and rationality that philosophers require of it.

> In particular, in order for an attitude to count as a belief, the attitude needs to be responsive to changes in the world, and in our evidential relation to it. (2008a: 563)

In this way, she is attempting to shut down arguments that she has invented an unnecessary new distinction, i.e., that we could already chalk up all these effects to *implicit* or tacit or unconscious belief.[17] The price of that move is that it limits the explanatory role(s) we need belief to play—"we need to save belief for more than these cases" (*ibid)*. Gendler also explicitly rejects a possible counterargument that the alief/belief distinction is just a renaming of the acceptance/belief distinction proposed by Cohen:

> Does alieving that P involve accepting that P? …Interestingly, the answer to this question turns out to be: no, and the way in which it turns out to be no reveals something important about the nature of alief. Unlike belief or pretense or imagination or supposition, alief does not involve acceptance. Though the point can be made on conceptual grounds alone, it is helpful to begin with a specific example... In [Rozin & Tuorila (1993) – described above], the label read precisely the opposite: it had "not sodium cyanide, not poison" written on it, with a red skull and cross bones preceded by the word not. So, although these subjects were in an alief state with the content "cyanide, dangerous, avoid," the content they were prompted to imagine was exactly the opposite. They did not—as the acceptance condition requires—regard it as true in some way that cyanide is to be found in the vicinity; instead, it was the negated presence of the word "cyanide" that rendered occurrent their cyanide-associated aliefs. (Gendler 2008b: 648-649)

The upshot of positing something like alief is that we gain a fairly elegant explanatory account of the many cases under discussion in which *belief perseverance*, *misinformation effects*, and *inconsistency awareness* failures complicate belief revision.

> Given the nature of alief and belief, it is inevitable that there will be cases where alief-generated propensities and belief-generated propensities activate contrary behavioral repertoires. The reason is simple: Aliefs involve habitual responses to apparent actual stimuli, but things may not be as they seem, the world may change, and one's norms may demand that the way things are is not the way things ought to be. Aliefs by their nature are insensitive to the possibility that appearances may misrepresent reality, and are unable to keep pace with variation in the world or with norm-world discrepancies. (Gendler, 2008a: 570).[18]

> If alief drives behavior in belief-discordant cases, it is likely that it drives behavior in belief-concordant cases as well. Belief plays an important role in the ultimate regulation of behavior. But it plays a far smaller role in moment-by-moment management than philosophical tradition has tended to stress (Gendler, 2008b: 663).

I will not defend a position on whether Gendler's proposition to add *alief* to our mental state lexicon is, at the end of the day, something to support or reject. I only note it as yet one more theory that has been put forward to help understand the apparent gulf between normative and

descriptive accounts of rationality. Also, her ideas will return in the discussion of self-deception and delusion in PART III, as in her (2007) she makes use of the concept in that regard. For now, I will move on to another account that suggests belief and believing is some form of dual process affair—namely, the account of psychologist Dan Gilbert.

1.3.3 – The 'Spinozan' and 'Cartesian'

Dan Gilbert is a social psychologist, who, somewhat interestingly, has developed a theory of "How Mental Systems Believe" (1991) that is predicated on a comparison of philosophical systems—namely, those of Descartes and Spinoza. On Gilbert's reading, Descartes (1641) is committed to the assumption that we must clearly comprehend a proposition in order to evaluate it for truth, and subsequently determine whether or not to believe it. Spinoza, on the other hand, contends that we cannot hold a proposition in mind without *first believing it*: the act of comprehension implies a provisional sort of belief, which can then be revoked upon active reflection.

> According to Spinoza, the act of understanding is the act of believing. As such, people are incapable of withholding their acceptance of that which they understand. They may indeed change their minds after accepting the assertions they comprehend, but they cannot stop their minds from being changed by contact with those assertions. Acceptance then, may be a passive and inevitable act, whereas rejection may be an active operation. (Gilbert, Tafarodi, & Malone, 1993: 222)

According to Gilbert, on the *Cartesian* model, a proposition must be understood *first*, then evaluated, and subsequently accepted as a belief, or rejected (and tagged "disbelieve"). On the Spinozan model described above, one first passively believes (in order to understand) and then proceeds to evaluate and can reject if necessary.[19] Gilbert's interest in modern philosophy isn't merely academic, however: his claim is that multiple recent findings on *belief perseverance* and *misinformation effects* in social psychology vindicate the Spinozan model.

Gilbert uses the metaphor of a library filing system to explain the distinction between the two sorts of system (Gilbert, Krull, Malone, 1990: 602). Imagine a librarian is tasked with sorting books in a way that distinguishes fiction from non-fiction. In one system—the *Cartesian*—each book is tagged with a red sticker, if fiction; or a blue sticker if it is non-fiction. In another system—the *Spinozan*—only fiction books are tagged with stickers, non-fiction books are left untagged. On the one hand, the Spinozan librarian here has it easier:

there is only one thing to look for (fiction), and fewer stickers need to be used. However, despite some small efficiencies, the Spinozan system has a problem that if, for some reason, a book makes it on to the shelf without having been checked for *fiction* status, it is effectively "labeled" *non-fiction*. The default book label is *non-fiction* in the Spinozan system. Gilbert argues that a Spinozan system of belief evaluation is similar: all ideas are by default tagged as "true", for to be able to understand an idea, one takes it to be true. As a result, certain ideas will metaphorically find their way on to the shelf "labeled" as true, since the mere act of apprehending them presents them as such—and without a conscious, active effort to subsequently evaluate the idea, the opportunity to tag it as *false*, and thereby to disbelieve it, can be complicated or even lost. And, just like the case of the librarian, an even small distraction could be enough to disrupt the processing in a way that allows false "non-fictions" to slip through.

> The most basic prediction of this model is that when some event prevents a person from "undoing" his or her initial acceptance, then he or she should continue to believe the assertion, even when it is patently false…These active measures require cognitive work (i.e., the search for or generation of contravening evidence). (Gilbert, Tafarodi, & Malone, 1993: 222)

Well, one might think at this point, all the more reason to reject the Spinozan account, and stick with a Cartesian one. However, Gilbert highlights numerous studies using a "misinformation debriefing" paradigm[20] (as discussed above, by Harman), including some he and his colleagues have published, which offer fairly convincing support that the human mind is indeed a Spinozan system that runs separate tracks for believing and unbelieving—the former being easy and unconscious, the latter much more difficult. In Gilbert, Krull, & Malone's 1990 study, they demonstrated that "interruption after comprehension leaves people in their initial state of acceptance, and that this state truly constitutes a belief insomuch as people will base consequential social behaviour on it" (Gilbert *et al.,* 1993: 223). The study used a jury paradigm, in which subjects were given crime reports in which some information had been color-coded and marked "untrue/disregard".

> The first report described how a perpetrator named Tom had robbed a stranger who had given him a ride, and the second report described how a perpetrator named Kevin had robbed a convenience store. Each report contained seven false statements that were printed in red. In one report, the false statements would have exacerbated the severity of the crime had they been true, and in the other report the false statements would have extenuated the severity of the crime had they been true... Some subjects saw a report whose false statements extenuated Tom's crime (described in the first

> report) and exacerbated Kevin's (described in the second report), and the remaining subjects saw a report whose false statements [did the reverse]. The false statements were constructed such that their elimination did not impair the grammatical integrity of the sentences in which they were embedded or the structural integrity of the crime stories themselves. In addition, the false statements were logically independent both of each other (i.e., the content of one neither implied nor refuted the content of another) and of the true statements (i.e., the content of a true statement neither implied nor refuted the content of a false statement or vice versa). We expected these false statements to affect the prison terms recommended by subjects who performed the digit-search task (the interrupted condition), but not those recommended by subjects who performed no digit-search task (uninterrupted condition). (1993: 224)

What they found is precisely as predicted: interrupted subjects were more likely than uninterrupted subjects to misremember false statements as true. However, both interrupted and uninterrupted subjects were equally likely to misremember true statements as false, which suggests the interruption was not impairing memory generally. "Finally, and most important, the number of false statements that subjects misremembered as true was reliably correlated with the length of the prison term they recommended" (*ibid*). This last point is one that has serious practical significance, given the numerous points in any trial where jurors may be asked to "disregard" something they just heard! Evidence suggests that the disregard instruction, even in cases where it is understood fully, and where jurors consciously believe they have successfully disregarded as instructed, the false information continues to infect and motivate judgment.[21]

The upshot of these and similar findings, according to Gilbert, is that "people are unable to decouple acceptance and comprehension, even when it would be propitious to do so" (Gilbert, 1991: 115). This, if correct, seems to throw a pretty hefty wrench into views on belief revision that involve decoupling acceptance from believing: on Gilbert's view, acceptance is automatic, rather than a *choice*, in the way a premising policy might be construed. But this automaticity of acceptance will make it that much more likely that a merely accepted premise could be mistaken for a belief, without having been through the proper evaluative steps. In short, we are generally (passively) *credulous*, and it takes effort to put credulity aside in order to evaluate and actively negate a proposition that has already been tagged "believe"'. Gilbert notes that this thesis fits well with research in linguistics (Bloom, 1970; Pea 1980) demonstrating that:

> the ability to deny propositions (i.e., truth-functional negation) is, in fact, one of the last linguistic abilities to emerge in childhood... Children are especially credulous, especially gullible, especially prone toward acceptance and belief—as if they accepted

> as effortlessly as they comprehended, but had yet to master the intricacies of doubt. In short, human children do precisely what one would expect of immature Spinozan (but not Cartesian) systems. (Gilbert, 1991: 110-111)

Additionally, he cites Horn (1989) whose studies have shown conclusive evidence that negation is a "second order affirmation: negative statements are about positive statements, while affirmatives are directly about the world" (Horn, 1989:3). Clark & Clark (1977) similarly have argued that linguistic evidence points to the thesis that people generally approach propositions/representations with the truth index set to *true,* and then proceed to compare with previously held *truths:* if they match, no change, if there is a mismatch, one must be switched to *false.* The reason for this, one might speculate, is that generally a charity principle is invoked: if the information being represented is linguistic in form, Gricean maxims will set the default evaluation to *true,* barring evidence of flouting maxims of quality, or relevance (Grice, 1975; 1989). If the information being represented is simply perceptual, again the default setting is set to *true*, as perception is generally accepted as veridical, barring evidence to the contrary

1.3.4 Dual process theories

So-called 'dual process' views of human rationality have proliferated since the pioneering work of Tversky & Kahneman in the field of "behavioral economics" beginning in the 1970s.[22] There are small variations in the dual process accounts that have been developed over those years, though the fundamental premise shared between them is that human cognition runs on two separate tracks, or *systems*: System 1 is characterized as reflexive, automatic, computationally frugal, associative, un- or sub-conscious, fast, skilled, intuitive, and evolutionary ancient (i.e., shared by animals); System 2, on the other hand, is reflective, controllable, computationally demanding, inferential, conscious, slow, deliberative, rule-governed, and evolutionarily recent (i.e., possibly restricted to *homo sapiens*).[23] The idea here is that System 1 handles "thinking" that a creature doesn't really have *time* to think about. Reflexes would be a good example: when the peripheral visual system senses a large projectile moving on an impact vector towards one's head, one *ducks*—it is not a conscious decision: indeed, only after it happens does the conscious mind seem to "catch up". A simple and intuitively plausible evolutionary story can be told to explain the existence of such an unconscious, lightning-fast cognitive subroutine: it saves the life of the organism efficiently

and at very little cost (i.e., the cost of a few "false positives" is basically nil: ducking when you didn't need to, such as while watching a 3D film, isn't going to hurt, whereas not ducking in a 3D world when you *need to* might hurt a lot)—hence, selective pressures favour creatures with such a cognitive subroutine.

Evans & Over (1996) suggest the fundamental difference between the two systems is their domain of operation: System 1 is a suite of "domain-specific" pragmatic operations that run beneath conscious awareness, and are tuned by and automatically responsive to the environment. System 2, on the contrary, is "domain-general" in the sense that it can operate according to rules and normative logical conventions across any domain towards which conscious attention is turned. System 2, as a result of generality and flexibility, is far more cognitively demanding and slower. Sloman (1996) describes the two systems similarly, referring to System 1 as an "intuitive processor", and System 2 as a "conscious rule interpreter." Sloman also discusses the ways in which the two systems can interface, introducing what he calls "Criterion S":

> A reasoning problem satisfies Criterion S if it causes people to believe two contradictory responses simultaneously. By 'believe', I mean a propensity, feeling, or conviction that a response is appropriate even if it is not strong enough to be acted on. (Sloman, 1996: 384)

His point here mirrors some of the findings we already discussed regarding the (sometimes) dissociation between our actions and our occurrent judgments—cases, for example, where we act in contradiction to our beliefs, or when we (unconsciously) allow putatively "discarded" (mis)beliefs to continue to influence thoughts and actions. We can *try* to consciously override System 1 using System 2, but will often fail, as System 1 processes will continue to run automatically, when in salient contexts.

Stanovich & West (2000) have perhaps the most fully elaborated dual process view, incorporating many of the "heuristics and biases" that have been empirically identified in social psychology, and attempting to offer an "interpretation of the gap between descriptive models and normative models in the human reasoning and decision making literature" (2000: 645). On their view, common reasoning biases can be chalked up largely to what they term the "*fundamental computational bias*"—which is the tendency to automatically contextualize problems (and hence engage System 1, even where System 2 would be more appropriate). In such cases, even under conscious reflection, using System 2, we may output judgments (or

behavior) that seems "irrational" in the sense that it is tainted by System 1 processes that are inconsistent with the global situation we are in.[24] Part of the problem, according to Stanovich & West, is language-based: Gricean conversational maxims (themselves an arguably System 1-mediated process: unconscious and automatic) will drive toward contextualization in any problem-setting where the language system is engaged (even if the problem is better suited to a more abstract, exclusively System 2 reasoning process).[25]

> Construals triggered by System 1 are highly contextualized, personalized and socialized. They are driven by considerations of relevance and are aimed at inferring intentionality by the use of conversational implicature even in situations that are devoid of conversational features. The primacy of these mechanisms leads to what has been termed the fundamental computational bias in human cognition—the tendency toward automatic contextualization of problems. In contrast, Systems 2's more controlled processes serve to decontextualize and depersonalize problems. This system is more adept at representing in terms of rules and underlying principles. It can deal with problems without social content and is not dominated by the goal of attributing intentionality nor by the search for conversational relevance. (2000: 658-59)

A classic example would be Tversky & Kahneman's (1983) famous study on the conjunction fallacy involving *Linda, the feminist bank teller*. In that study, subjects are presented with a vignette of Linda, a young woman described as fitting a largely "progressive" stereotype, and proceed to allow that stereotype to (mis)lead them to a clearly illogical judgment. In the study, subjects were given the following vignette and instructions:

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.*
>
> TASK: Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable:
>
> a. Linda is a teacher in an elementary school
> b. Linda works in a bookstore and takes Yoga classes
> c. Linda is active in the feminist movement
> d. Linda is a psychiatric social worker
> e. Linda is a member of the League of Women Voters
> f. Linda is a bank teller
> g. Linda is an insurance salesperson
> h. Linda is a bank teller and is active in the feminist movement

The key finding in this study is that a majority of subjects fall prey to the "conjunction fallacy" and conclude that *h* is more probable than *f*, despite this being clearly illogical as *h* is a proper subset of *f*.[26]

Frankish (2006) brings the dual process view to bear specifically on *belief*, arguing for a "two-strand" account of belief:

> one which is conscious, flat-out, capable of being actively formed, often language-involving, and, consequently unique to humans and other language-users; and another which in non-conscious, possibly not subject to occurrent activation, partial, passively formed, probably nonverbal, and common to both humans and animals. (Frankish, 2004: 22)

The two sorts of belief, which Frankish names *basic beliefs* and *superbeliefs,* coincide roughly with System 1 and System 2 processing, respectively, as discussed above. The new descriptor in Frankish's account is "flat-out" with respect to *superbelief*—by this he means that such beliefs, on account of being the product of System 2 processes, respecting logical norms of operations, will terminate in discrete, non-graded belief states, corresponding to attitudes of *truth* or *falsity* toward the proposition in question. The *basic beliefs* that are the product of System 1, on the other hand, are not flat-out, as they are the product of (unconscious) Bayesian procedures, which output only probabilities.[27]

Frankish employs the belief/acceptance distinction (citing Cohen, 1992) as a way to differentiate the two strands of belief, in the sense that System 2 works via "premising policies" (Frankish, 2004: 84). Connecting System 2 *superbelief* to "acceptance", according to Frankish, can help explain many of the rational "breakdowns" that have been highlighted so far throughout this chapter (contradictory belief sets, acting in opposition to professed belief, biased reasoning, etc.), as he suggests we view *beliefs* as merely "unrestricted acceptances" (135). He concludes:

> On the view outlined here, an action can be assessed for rationality at two different levels and in accordance with two different sets of norms: at the supermental level in accordance with the norms of classical practical reasoning, and at the basic level, in accordance with the norms of Bayesian decision-making... In working out the consequences of a set of premises and goals [at the supermental level] we may go astray. We may misapply an inference rule, or apply an invalid one, or assume that the response to a self-interrogation is correct when in fact it is not. Indeed, we may go wrong blatantly and systematically. And if we act upon the conclusion of such faulty reasoning, then the resulting actions *qua supermental*, will be irrational. (2004: 147)

1.4 Review and look ahead

Let's take stock for a moment, and see what 'dual process' theory buys us in terms of central topic in this dissertation: belief revision. On the one hand, if there are two distinct levels of processing, one essentially automated, fast and unconscious, and the other controlled, slow

and conscious, then we can start to get a clear picture as to *why* normative accounts of belief revision seem detached from actual belief revision failures. *Inconsistency awareness* is going to be a problem, if much of our "reasoning" is happening via System 1. To become *aware* of an inconsistency is going to be, *a fortiori,* a System 2 process. However, as we have seen, System 1, triggered by contextual cues, will often interfere. Sometimes we will still be made aware of an inconsistency *after* the fact (e.g., if we commit the conjunction fallacy in the Linda/bank teller case, we can realize the mistake *after*, upon conscious re-evaluation)—but this won't change the fact that in a similar problem context later, we will almost certainly *make the same mistake again.* So we can explain the *inconsistency awareness* and perhaps *belief perseverance* issues quite well using dual process theory. However, we will still be left with a glaring problem concerning *framing* and cognitive resources: System 1 is predicated on the conservation of cognitive resources, with dedicated subroutines optimized for dedicated domain-specific problems. But System 2 is domain-general, and expected to be able to run processing *over the entire belief set*. This explains its "slowness," perhaps—but, as we will discuss in the next section, "slow" doesn't come close to describing the computational strain that global access implies. System 2 still encounters the *finitary predicament* and what we will elaborate below as the *frame problem.*

Notes for chapter 1

[1] Some write it "Quineian"—like Fodor, for instance, who will be the subject of much discussion below. I prefer without the "i". But "Quinean" and "Quineian" should be taken to mean the same thing.

[2] As we shall see, through numerous examples in this dissertation, we often know better, and fail to do better.

[3] Note that motivational biasing to form and retain beliefs that one has no or little (or even disconfirmatory) evidence for is an issue that will be discussed at length in chapter 6. Again, as in the case of *inconsistency awareness*, this suggestion that one inspect one's biases and motivations for believing any given proposition presupposes a transparency and availability for introspection that may be illusory.

[4] This makes all evidence observational, which again implies a transparency to the process: "Observations stubbornly retain their primacy. They remain the boundary conditions of our body of beliefs" (31-32).

[5] Cherniak (1984; also Carruthers, 2006b, 2006c) uses a similar example to argue against the idea that we actually check individual beliefs for consistency. Of course, this may seem like a strange and laborious way to check for consistency: why would we check them all anyhow? We already believe the original 150 to be TRUE. Cherniak's point is simply that checking for tautological consequence and truth-functional consistency is certainly NOT the way we go about it—it cannot be, for computational reasons.

[6] I speak of a "standard" model of abductive inference here, for the sake of simplicity and brevity. As Lipton (2004) notes: "Inference to the Best Explanation has become extremely popular in philosophical circles, discussed by many and endorsed without discussion by many more... Yet it still remains more of a slogan than an articulated account of induction" (Lipton, 2004: 57).

[7] Peirce similarly describes the process of abductive inference as somewhat opaque, and more akin to "insight":

> The abductive suggestion comes to us like a flash. It is an act of ***insight,*** although of extremely fallible insight. It is true that the different elements of the hypothesis were in our minds before; but it is the idea of putting together what we had never before dreamed of putting together which flashes the new suggestion before our contemplation. (Peirce, 1931: §5.180)

[8] The reference to 'satisficing' as the deflationary criterion is from Simon (1956) – we will return to this idea again in §2.1, below.

[9] Harman recognizes that the inconsistency avoidance principle is roughly correct, but "defeasible":

> like the case in which an author believes each of the things he or she says in a book he or she has written and also believes, given human fallibility, at least one of the things he or she has said in the book must be false. Such a person is justified in having inconsistent beliefs, but that does not show that the Recognized Inconsistency Principle is incorrect. It only shows that the principle is defeasible. (1986:24)

[10] Note that Harman cites Anderson's (1982) discussion of the findings in Anderson & Ross (1975).

[11] E.g., Anderson, Lepper, & Ross (1980); Anderson & Schacter (1986); Johnson & Seifert (1994); Ross, Lepper, & Hubbard (1975); Lord, Ross, & Lepper (1979); Lord, Lepper & Preston (1984); Schul & Burnstein (1985); Wilkes & Leatherbarrow (1988); Wegner et al (1990).

[12] Interestingly, Harman proposes a relatively minor condition to be placed on belief revision that, given what we have just said, could pose *major* trouble down the line. His *get back principle* states that one should not give up a belief that will be easily "gotten back"—e.g., in such cases where "one can usually get back the dropped belief by reviewing the reasoning that led to it in the first place" (Harman, 1986: 58). Harman means this in the sense that other beliefs in the system will regenerate the rejected belief via implication. This will be interesting when we discuss subdoxastic or unconscious pathways to belief in later sections of this dissertation. If we should *not* give up beliefs that are easy to "get back" merely by reviewing what led to them—as Harman suggests—then we may have a problem if some of those beliefs can be "gotten back" immediately, not via conscious (re)-reasoning, but by mere experience, or inferential reasoning processes which may be occurring beneath the surface, and inaccessible to conscious reflection.

[13] Frankish (2004: pp. 126-128), discussed at more length below, is helpful for elucidating these distinctions—it is from Frankish that I borrow the mentioned example of the lawyer who "accepts" her client's innocence as a premise while preparing a defense yet surely need not, and may actually not, *believe* it.

[14] She references Hume who similarly discusses the interesting case of the caged men who "cannot forebear trembling ... tho' he knows himself to be perfectly secure from falling, by his experience of the solidity of the iron, which supports him" (1739/1978:148; Gendler 2008a: 553).

[15] Note that the skydeck actually cracked under some tourists' feet just recently (Lutz, 2014). So perhaps trembling is relatively reasonable after all!

[16] We'll come back to this study when we look at "ironic process theory" in Chapter 7. This may be a textbook case of that phenomenon: the proposition that the water in front of one is *not* poisoned triggers also the thought of poison, and hence the associated *alief*—'poison, bad for me, don't drink'.

[17] "Tacit" belief is the term Lycan (1986) uses. I will prefer "implicit" in this dissertation, given the connections that will be made between belief and memory research in chapter 7, below. In the memory literature, the distinction is made between *implicit/explicit*—though, admittedly, the definitions of these in the cognitive science literature do not map perfectly onto the distinction Lycan and other philosophers draw between tacit/occurent belief. I will use "implicit" to mean roughly *unconscious*, *beneath the level of immediate awareness.*

[18] Schwitzgebel (2010: 539) argues that Gendler has "overdrawn the distinction" here, arguing that "[o]ur habits, associations, and automatic responses *are,* to a substantial extent, responsive to evidence; and our verbal avowals or dispositions to judge are often *un*-responsive to evidence"—contrary to Gendler's claim that alief is non-responsive to evidential defeat, while belief must be. To some extent, I sympathize with Schwitzgebel's critique: take as an example a common situation that Frankish (2004) brings up: when you know the light bulb in the kitchen is burned out, but you nevertheless flick the switch when you enter the room (apparently immune to your own knowledge)—this seems like a good candidate for *alief-hood.* But it *is remediable*.

[19] In this discussion of Gilbert's view, I will accept his reading of Spinoza (and Descartes) without dispute. A full exegesis of Spinoza's view on the matter would take us too far afield.

[20] We will discuss a number of these at more length, below, in the section on *intentional forgetting*. In this section, we will look just quickly at Gilbert's own findings. Similar studies include Nisbett & Wilson, 1977; Arkes, Boehm & Xu, 1991; Arkes, Hacket & Boehm 1989; Begg, Armour & Kerr 1985; Hasher, Goldstein & Toppino 1977; Anderson 1982; Lindsay 1990; Ross, Lepper & Hubbard 1975; Schul & Burnstein 1985; Wilson & Brekke 1992; Wyer and Budesheim 1987; Wyer & Unverzagt 1985; Wegner Coulton Wenzlaff 1985.

[21] See Pickel's (1995) "Inducing Jurors to Disregard Inadmissible Evidence: A Legal Explanation Does Not Help" for a complete rundown of studies which highlight this effect. More on this in chapter 7.

[22] What's also known as the "heuristics and biases" program. Much more will follow on this in multiple sections below, especially in chapter 6, where we will highlight many of the findings from this research program.

[23] See Osman (2004) for a thorough run-down of dual process theories.

[24] Yet another failure of *inconsistency awareness*.

[25] Stanovich & West differ from some other dual process theorists insofar as they are "meliorists" and believe that the fundamental computational bias is to some degree remediable, and the descriptive/normative gap may be reducible—we can retrain our cognitive systems to better anticipate System 1 interference in System 2 reasoning task in order to avoid bias—we simply need to study and delineate the sorts of contexts in which System 1 interference is likely. As we will see in chapter 8, this may have some relevance for clinicians seeking to help delusional patients overcome and/or change behavior patterns based on stubborn false belief.

[26] This is an example of the *representativeness* heuristic, as defined by Tversky & Kahneman (1973). In chapter 6, we will return to the *heuristics & biases* research program, and discuss examples such as these at greater length.

[27] It would take us too far afield here to elaborate extensively on Bayesian inference principles—in the most basic terms, Bayes Theorem (Bayes & Price, 1763) formalizes conditional probability estimation—how to update belief (or hypotheses) probabilistically, based on prior knowledge (or evidence). The formula is $P(A|B)=[P(B|A)P(A])/P(B)$, where $P(A|B)$ is the *posterior*, $P(B|A)$ is the *likelihood* (or base rate), $P(A)$ is the *prior,* and $P(B)$ is the *evidence.* Note that Lipton (2004) says abductive inference is Bayesian – so that may not fit with dual systems approach, as Frankish envisages it. (I.e., Frankish seems to argue that Bayesian inference is an unconscious system 1 affair – Lipton describes abductive inference as essentially Bayesian, but also clearly a *conscious, deliberative* activity. So they are suggesting we employ Bayesian reasoning in very different ways.)

2 The 'finitary predicament'

In this chapter, we will look at two influential accounts of rationality that attempt to incorporate empirical findings from cognitive science regarding processing limitations—what Cherniak (1986) calls the "finitary predicament" of human reasoning—that will (or should, at any rate) circumscribe any comprehensive theory of belief revision. I will look first at Cherniak's (1986) *Minimal Rationality*, and then at the work of Gigerenzer & Todd, who in many papers jointly and severally have argued for a position they call "ecological rationality" (Gigerenzer & Todd, 1999; Todd & Gigerenzer, 1999; Gigerenzer 2011)—built on the idea of "bounded rationality" (Simon, 1962), and informed by numerous studies and findings of psychologists, economists and philosophers contributing to the "heuristics and biases" literature and the development of "dual process" theories discussed in chapter 1. After reviewing the arguments based on the "finitary predicament" we will then turn to the *frame problem*, which arises as a direct result, and poses serious obstacles to any comprehensive account of belief revision (or cognition in general, as we will see). Proposed "solutions" to the frame problem will be explored in PART II of this dissertation, which is comprised of a lengthy argument in favour of cognitive modularity, from bottom to top.

2.1 Minimal rationality

Cherniak's *Minimal Rationality* takes aim squarely at accounts, like those discussed above in §1.2, that to varying degrees prescribe belief revision practices that make demands on our cognitive practices that can't, in actuality, be met, given architectural constraints. He argues:

> one cannot even explain important ranges of actual human behaviour without employing models more "psychologically realistic" than conventional philosophical ones [...] The belief systems of human beings do not inevitably and automatically readjust themselves appropriately in the way Quine describes. (Cherniak, 1986: 49-50)

The "way" of belief revision he is attributing to Quine here is the *ideal consistency condition:* "If A has a particular belief-desire set, then if any inconsistency arose in the belief set, A would eliminate it" (17). Granted, as we saw above, Quine recognizes that often contradictory beliefs will persist due to ignorance of the contradiction—however, his account

does not *excuse* the misbeliever on that ground, but rather simply explains it. I have already suggested, and we will pursue the argument at length later as well, that the problem of persistent, perhaps unavoidable, *inconsistency awareness* dooms the standard normative account of belief revision to being essentially useless in practical terms. Cherniak's complaint is somewhat different: he argues that the *ideal consistency condition* is plagued, not by unawareness of inconsistency, but by the combinatorial explosion of cognitive resources that ideal consistency demands.

> The most important unsatisfactoriness of the ideal general rationality condition arises from its denial of a fundamental feature of human existence, that human beings are in the *finitary predicament* of having fixed limits on their cognitive capacities and the time available to them. (8)

Cherniak's proposal is a more *minimal consistency condition:* "If A has a particular belief-desire set, then if any inconsistencies arose in the belief set, A would sometimes eliminate some of them" (16). The idea here would be to recognize the intuitively obvious point that inconsistencies are to be *avoided*—that if we couldn't expect at least a minimal level of inconsistency-avoidance, then the attribution of belief "could not be of any value in predicting the agent's behaviour" (*ibid)*—while resisting the idea that failure to root out inconsistent beliefs would amount to either a) patent irrationality; or b) not having beliefs, properly so-called.

Now, if we are expecting to *some*times eliminate *some* inconsistencies in the belief set, we will of course likely want to set some range of *reasonableness* on how minimal the expectation is. As we will see in later chapters, below, attributions of rationality (and of belief possession) are generally relativized, at least in part, to the number and severity of inconsistent or contradictory beliefs and belief-desire pairs. Cherniak does not directly address the boundary conditions on this point. I will elaborate on my own attempts to do so in the final sections of this dissertation: for now, in the interest of laying down a marker on my own final account, I think that the *minimal inconsistency condition* will end up met by pretty much everyone, even agents considered to have 'pathological' belief states/systems, and those classed as unambiguously and floridly delusional.[1] The upshot of this would be that perhaps Cherniak's *minimal inconsistency condition* does not clearly demarcate the line between what we commonly refer to as "rational" and "irrational" in the way he intends it to. (Much more on this later).

To return to Cherniak for a few final points, his account has much to offer in terms of supporting the thesis that minimal conditions on rationality are appropriate given cognitive resources, and diagnosing the reasons why philosophical accounts of belief revision have tended to be blind to this reality.[2] Cherniak locates the problem with standard philosophical accounts in the failure to respect the distinction between short and long term memory, and the computational organization of storage, access, and transfer between the two systems.

> Failure to acknowledge the long-term/short-term memory distinction seems responsible for most of the denials common in philosophy of the possibility of people making obvious logical errors. (56)
> [...]
> The short-term/long-term memory distinction entails that only beliefs in short-term memory can be premises in reasoning; beliefs in long-term memory are inert—they do not interact with each other, and they do not affect behaviour. (59)

If Cherniak is correct that only items in short term memory[3] can serve as premises in reasoning, then we have a clear explanation of how people fail in the way philosophers suggest they should not, such as holding contradictory beliefs, violating transitivity in preference formation and valuing, failing to believe the immediate implications of one's beliefs, etc. In chapters 5-7, we will look more closely at the data that supports his contention—for now, we will grant this point (I do, in fact, agree that this is correct), and note that this fact about memory seriously limits how much of the belief set can be inspected (and reasoned with) at any given time. The times in which we will become aware of inconsistencies will be limited to the times when both the beliefs p not $\neg p$ are synchronically available in short term or working memory.

In order for memories to be effectively stored so that they can be retrieved as needed for 'on-line' processing (reasoning), Cherniak suggests that the only plausible architectural design that could subserve this need is one where

> the contents of long-term memory are organized. An item in long-term memory is located for retrieval not by a search of the entire memory but by a narrower search that takes advantage of the structure of the memory. All of these accounts in effect represent the long-term memory as a graph-theoretic entity, a network of nodes interconnected by arcs. The model is a generalization of a filing system where a file can in turn contain subfiles. (53)[4]

The reason for insisting that memory is structured in this way is that searches of memory have to be both *tractable*, insofar as they need to be able to focus on appropriate information given time constraints; and also *effective*, in the sense that they deliver the appropriate beliefs

up to conscious deliberation. Cherniak references Hume (1739/1978) regarding the mere fact that we are capable of doing this—of searching only *part* of our memory, yet coming up with exactly what we need, generally—seems somewhat "magical".[5] Cherniak suggests that it isn't magic, we must simply accept that evolutionary pressures have shaped our minds to store information in "compartmentalized" ways that allow for *non*-exhaustive, yet relevant and useful searches.

> If there is no compartmentalization, if there is an equal likelihood that any belief will be recalled in conjunction with any other belief, cognitive resources will be spread too thin. Some degree of compartmentalization is indispensable for adequate management of our large memories [...] We now have the solution to Hume's mystery of how partial memory search procedures can be adequate; no magical homunculus is necessary. We have found a connection between memory organization and rationality: a basic precondition for our minimal rationality is efficient recall, which itself requires incomplete search, which in turn requires compartmentalization. (68-69)

Cherniak's insistence on the necessity of "compartmentalization" is certainly well-motivated here: without compartmentalization—without a system of tagging and filing beliefs in memory, including associative links *between* beliefs, which in turn can be the subject of further (higher-order) belief—searching would be blind and slow. Successful compartmentalization can streamline the search process, conserve cognitive resources, and respect the "finitary predicament".[6]

However, even with this idea in mind, it seems we will still bump into another aspect of the "finitary predicament" that we mentioned in passing in §1.2.2—the *frame problem*, or the problem of determining relevance. On Cherniak's account, our minds are simply well-adapted in the sense that memory storage is compartmentalized in a way that facilitates quick and effective retrieval (even though this comes at the price of incomplete consistency-checking in many instances). However, the problem of relevance-determination pops up on both ends of this process: on the one hand, how does the system "decode" the current deliberative context in a way that directs searches to the right compartments? And on the other hand, perhaps even more puzzling, when memories are tagged for storage, doesn't this imply some sort of executive function that can oversee *the whole system*, including all of its compartments? That sort of executive control seems impossible for the very reasons Cherniak has argued for compartmentalization in the first place. We are getting closer and closer to the full-blooded *frame problem*, which will be the focus of §2.3, below. But first,

it's worth looking briefly at another "minimalist" account of rationality and belief revision: the account of *bounded* or *ecological rationality.*

2.2 Bounded rationality

The term "bounded rationality" was initially coined by Herbert Simon (1956; 1982) to dissociate real-world human rationality from the normative models discussed generally in philosophy that condition rationality on seemingly unbounded search and inference strategies that, referencing Laplace (1814), assume "demon"-like capacities.

> Humans and animals make inferences about their world with limited time, knowledge, and computational power. In contrast, many models of rational inference view the mind as if it were a supernatural being possessing demonic powers of reason, boundless knowledge, and all of eternity with which to make decisions. Such visions of rationality often conflict with reality. (Gigerenzer &Todd:1999, 5)

Simon (1956) argues that optimization is not feasible, so we must invoke a "satisficing" criteria instead.[7] As Gigerenzer & Todd explain it:

> Satisficing is a method for making a choice from a set of alternatives encountered sequentially when one does not know much about the possibilities ahead of time. In such situations, there may be no optimal solution for when to stop searching for further alternatives—for instance, once Darwin decided to marry, there would be no optimal way of deciding when to stop looking for prospective marriage partners and settle down with a particular one. Satisficing takes the shortcut of setting an adjustable aspiration level and ending the search for alternatives as soon as one is encountered that exceeds the aspiration level. (G&T, 1999: 13)

Gigerenzer & Todd incorporate Simonian "satisficing" into their own version of *bounded rationality,* which envisages "a *bounded* mind reaching into an adaptive toolbox filled with fast and frugal heuristics" (G&T, 1999: 5). These "fast and frugal heuristics" are assumed to be essentially cognitive subroutines that have evolved due to adaptive pressures. Heuristics[8] are employed in deliberative tasks, in order to—as the name implies—conserve time and resources, while still delivering the desired result(s), or close approximations thereof—enough to *satisfice.* Gigerenzer & Todd refer to their revised bounded account as "ecological rationality": or "rationality that is defined by its fit with reality" (5).

The cognitive, deliberative heuristics posited by Gigerenzer & Todd can be divided into 3 main categories: those designed for 1) guiding searches; 2) stopping searches; and 3) decision-making (1996: 16-17).

> Fast and frugal heuristics employ a minimum of time, knowledge, and computation to

> make adaptive choices in real environments. They can be used to solve problems of sequential search through objects or options, as in satisficing. They can also be used to make choices between simultaneously available objects, where the search for information (in the form of cues, features, consequences, etc.) about the possible options must be limited, rather than the search for the options themselves. Fast and frugal heuristics limit their search of objects or information using easily computable stopping rules, and they make their choices with easily computable decision rules. (G&T, 1999: 14)

An example of a heuristic discussed by Gigerenzer & Todd is the *take the last* rule– a selection heuristic which directs an agent, under time pressure to make a choice, to (arbitrarily) select the same option as the time before—if that strategy worked the last time—as it may very well work again. One can tell a plausible evolutionary story for why such a rule would become entrenched in a cognitive system (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 another, untested choice.

In this section we have seen how traditional accounts of belief revision will ineluctably crash into the problem of limited cognitive resources, and we have explored various options to minimalize and/or bind accounts of rationality and belief revision to respect the *finitary predicament* of human cognition. Along the way, and in previous sections, we have been gesturing more frequently at a deep problem for accounts of belief revision: the *frame problem.* In the next section, I will elaborate on this problem at length, after which, in the following chapters, I will defend an account of cognitive architecture that potentially obviates the concerns about unboundedness and framing.

2.3 The frame problem

Historically, there have been a few iterations of the so-called *frame problem,* starting out as a simple representational problem in A.I. research, and morphing into a larger puzzle regarding belief revision, abductive inference, and deliberation in general.[9] I will focus on Fodor's (1987; 2000; 2008) iteration of the problem, which he calls "Hamlet's problem"—the problem 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). In order to best understand this, let's begin with a quick interlude regarding Hamlet to make the stakes clear.

2.3.1 To be(lieve) or not to be(lieve)

Arguably no single character in English literature has commanded so much attention and scrutiny as the melancholy Dane—he is a fascinating psychological study, a puzzle, whose 'breakdown' is amenable to various readings. On the dominant interpretation, Hamlet is pathologically indecisive—he thinks too much and too long—until "the native hue of resolution/ Is sicklied o'er with the pale cast of thought/And the enterprises of great pith and moment/With this regard their currents turn awry/And lose the name of action (Hamlet, III, i). 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 depression, self-loathing, and a quite literally *terminal* indecisiveness. There are so many questions that need to be answered, experiments to be run to test the reactions of others, 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 was 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.

Is this *rational* though? His decision procedures are certainly not optimal. But are they pathological? Is Hamlet actually *crazy?* Those around him certainly perceive his thoughts and actions as those of a madman. And part of his strategy to find the certain

knowledge and truths he requires him to *act* crazy, maybe even to *be* crazy.[10] To the audience, it's often an open question whether his madness is feigned or real—like Hamlet, we too suffer from an inability to determine how to separate reality from appearance—'being' from 'seeming'. Hamlet is paranoid in the extreme about the intentions of all around him—rightfully so with regard to Claudius, almost certainly mistakenly with regard to Ophelia, (and we're left uncertain whether his mother was worthy of his suspicion). He almost certainly would fit the DSM-IV diagnostic criteria for bipolar disorder, and possibly borderline personality disorder. But are his thoughts and actions *disordered?* Or are his 'breakdowns' exactly what one might expect a functional and rational person to experience under the stress and strain of extraordinary circumstance? As Hamlet himself proclaims, "I am but mad north-north-west. When the wind is southerly, I know a hawk from a handsaw" (Act II, scene ii).

Now jump ahead roughly 400 years and, with Hamlet in mind, enjoy a fable from 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 failed to reason through their existential dilemma in time to save themselves from destruction, just as Hamlet failed. 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. Or perhaps we just dismiss them as irrational—as crazy—though this is not likely an epithet that can be applied to the robot as easily as the melancholy Dane.[11] Either way, the dramatic irony in both stories is that we sense that *we* could judge the relevant issues, engage the correct beliefs and prescribe the appropriate mode of action to be taken. So why are these doomed protagonists so desperately unable to do the same?

The central 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. In one sense, this isn't really a problem—something in nature has already solved it—the fact is that we *do it* 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)

In disputes between various models of cognitive architecture, it's a charge leveled by nearly everyone: 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, resulting in combinatorial explosion. Cognitive processes, if they are going to plausibly respect tractability, require clear and efficient *halting* procedures—some way to frame the task at hand, impose frugality, and acquire and revise belief in a way that

avoids having to engage in exhaustive searches. Otherwise, we would all end up like Hamlet, and Dennett's R-series robots.

Of course, we are not, generally, all Hamlet—we do manage to revise belief, at least in some cases, efficiently and satisfactorily—so we are assured there *is* at least one sort of cognitive system that can work around the frame problem: namely, our own. The question is *how does it work?* How do we update and revise our beliefs? How do we come to start believing, and how to we come to cease believing, in the appropriate way(s)? And how does all of this take place in a computationally tractable fashion?

2.3.2 The Fodorian iteration 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 in cognitive science. 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*. Fodor suggests Pinker misses 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).[12] Fodor gets quite exercised about what he calls the "New Synthesis" school of cognitive science, as typified by Pinker (1997) and fellow travellers 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). Fodor 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 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).

Fodor is a self-professed "Quinean" on this point, arguing that not just abduction, but *all* inferential practice—along with analogical reasoning, scientific confirmation, and belief revision in general—is *isotropic.* He has been very firm on this point from his (1983) book on *Modularity of Mind*, where he highlights belief conformation in science—a prototypical abductive enterprise:

> Confirmation in science is isotropic and it is Quineian. It is notoriously hard to give anything approaching a rigorous account of what being isotropic and Quineian amounts to, but it is easy enough to convey the intuitions . By saying that confirmation is isotropic, I mean that the facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established empirical (or, of course, demonstrative) truths. Crudely: everything that the scientist knows is, in principle, relevant to determining what else he ought to believe. In principle, our botany constrains our astronomy, if only we could think of ways to make them connect. (1983: 105)

25 years later, in *LOT2:* The *Language of Thought Revisited,* Fodor is still banging the Quinean drum regarding belief revision:

> Typical nondemonstrative inference is *isotropic*. That's to say that, in principle, *any* of one's cognitive commitments (including, of course, the available experiential data) is relevant to the (dis)confirmation of any new belief. There is, in particular, no way to determine a priori what might turn out be germane to accepting or rejecting an empirical hypothesis. It is one of the differences between a theory of scientific *confirmation* (say, an inductive logic) and what positivists used to call a 'theory of

> scientific *discovery*' that the former simply takes for granted that both the hypotheses to be assessed and the data relevant to their assessment are specified *prior to* the computation of confirmation levels. It is then left to the theory of discovery to explain how the relevance of the data is to be estimated (and, for that matter, where the candidate hypotheses come from). (Fodor, 2008: 115)

Non-demonstrative inference, inference to the best explanation, scientific confirmation, all imply precisely the 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. But as Fodor laments, *everything* a scientist (or anyone) knows is, in principle, relevant to determining what else she ought to believe. That's a lot of things to check, and as discussed earlier, a seemingly *impossible* number of things to check—Quinean responsibilities notwithstanding.[13]

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—which sounds worryingly close to positing some sort of homunculus. Fodor is not arguing that there *is* such a homunculus, but he is suggesting that cognitive science, in particular the "new synthesis" school of computational cognitive science, is 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 chapter, we will look at one such area—modular processing—as it may hold a key to answering Fodor's frame problem regarding holistic reasoning.

2.4 Review and look ahead

The purpose has this chapter has been to highlight the central problem with standard accounts of belief revision—namely, what we are referring to as the 'frame problem'—how can a cognitive system, with limited computational resources, effect the sorts of context- and relevance-determining operations that are necessary preconditions to even the most basic or minimal belief revision procedures? As we have seen, updating belief in dynamic contexts, even on a very small scale, can be computationally challenging. Once we move beyond local updating to more global coherence checking and holistic belief management—i.e., what standard normative accounts of rationality demand of us—we run into an exponentially more difficult frame problem: how can we achieve even minimal *inconsistency awareness* without running constant exhaustive search procedures*?*

The answer to these questions will tell us a great deal about how the human cognitive system is (or must be) architecturally designed. As I will argue in the next few chapters, the only account plausibly in the running in this respect will be some sort of *modular* account. Accordingly, this will have repercussions on any normative account of belief revision and rationality, insofar as the practical constraints it entails. Additionally, a modular cognitive architecture will help account for numerous empirical findings regarding memory, belief perseverance, reasoning biases, implicit attitudes, and self-deception, all of which I explore in PART III. Finally, this modular account and the 'answer' to the frame problem that I will defend motivates a revision of current views on the nature of certain pathological belief states—namely, monothematic delusions—and offers both some novel ideas for a positive research path and the potential of more effective treatment strategies for psychologists and psychiatrists working on delusion.

Notes for chapter 2

[1] Cherniak's thesis presages some of what I will argue, as he contends that the explanatory price of a cognitive architecture that can successfully and efficiently function under minimal constraints essentially means the "global rationality requires some local irrationality" (1986: 70).

[2] Many of Cherniak's argument points will be returned to in chapters 5-7, when we turn to a discussion of empirical findings on memory.

[3] For now, we'll defer to Cherniak's terminology – though as will be seen in the chapters that follow, it is probably more correct to refer to *working memory* in this regard, following Baddeley (1986), rather than short

term. There are some terminological disputes and open questions regarding exactly how many levels of memory there are, and in what way conscious deliberation 'moves' memory from one system to the other. See Tulving (1991: 11-18) for a full discussion of historical disputes over how to subdivide the memory system(s). Regardless, Cherniak's main contention here is that in order to 'work' with a memory, you have to first retrieve it from long term storage.

[4] This description nearly perfectly envisages the account Fodor (1998, 2008) gives regarding concept storage and organization—and account that we will examine at length in chapter 5 below.

[5] Hume, from the *Treatise*, Book I, Part I, §vii:

> Nothing is more admirable, than the readiness, with which the imagination suggests its ideas, and presents them at the very instant, in which they become necessary or useful. The fancy runs from one end of the universe to the other in collecting those ideas, which belong to any subject. One would think the whole intellectual world of ideas was at once subjected to our view, and that we did nothing but pick out such as were most proper for our purpose. There may not, however, be any present, beside those very ideas, that are thus collected by a kind of magical faculty in the soul, which, though it be always most perfect in the greatest geniuses, and is properly what we call a genius, is however inexplicable by the utmost efforts of human understanding. (1739/1978: 24).

[6] This idea of "compartmentalization" will be echoed in the sections that follow in this dissertation, as we turn to a discussion of modular cognitive architecture and its (arguably) defining characteristic: *encapsulation.* Encapsulation goes beyond compartmentalization in that it organizes not only storage, but also processing algorithms themselves into discrete units that engage where appropriate, rather than simply *everywhere* in a way that courts combinatorial explosion. Much more on this in PART II of this dissertation.

[7] Note that we previously encountered this term in §1.2.2, as Harman references it.

[8] I should stress here that our discussion of "heuristics" here is focused exclusively on Gigerenzer & Todd's description and definition. There is a vast literature crossing economics, psychology, and rationality theory that discusses "heuristics" as well, though somewhat differently, mostly focused on the use of that term with reference to seminal studies by Tversky & Kahneman (1974; 1982; Kahneman & Tversky, 2000; Kahneman *et al.,* 1993; Kahneman & Miller, 1986; Gilovich, 1991; Gilovich, Griffin, Kahneman, 2002). We will return to a longer discussion of this latter "heuristics and biases" literature in chapter 6, below.

[9] Over the past 40 years, usage of the title 'the frame problem' has come to refer to somewhat different things, depending on the discipline in which it is being posed. Pylyshyn (1987) sketches 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: how to determine what axioms needed to be explicitly programmed 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 section is more in accord with the latter, wider reading of the issue as a computational one – "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" and that "Fodor doesn't know the frame problem from a bunch of bananas" (Hayes 1987: 132). Dennett (1987) 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 seems 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.

[10] Remember Cherniak's admonition, above, that "global rationality requires some local irrationality" (1986: 70).

[11] And, as I will argue in chapter 8, below, with regard to delusion in general: such behaviour is not well described as "irrational"—it is merely a misfire, or breakdown, of systems which subserve "rationality".

[12] 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).

[13] Fodor points out that most attempts to model this computationally utilize 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). "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)

PART II

3 Local modularity

Frame problems pop up immediately in cognitive processing long before we get the complicated questions of belief revision and norms of rationality. The mere task of representing the external world through sensory organs is fraught with computational challenges and a fundamental *frame problem* at the level of initial perceptual systems. We will take visual perception as a starting point—noting that everything we say in this regard does not hold just for humans, but for any organism that visually represents its environment. Trying to understand the mechanisms of visual perception presents us with what, in mathematics, is referred to as an *ill-posed problem*—namely, one whose solution is either a) non-existent, b) not unique, and/or c) not continuously dependent on the initial data (i.e., it is highly sensitive to small shifts or "noise" in the data).[1] Poggio (1985) notes that the ill-posed nature of the problem of vision shows up at the earliest stages: the effort to represent a 3D world via two-dimensional retinal image presents serious computational difficulties, and hence a frame problem, on a number of levels. For example, a curve represented in two dimensions could be the result of infinitely many curves in three dimensions—think, e.g., of a circular tabletop viewed from an angle: it will appear as an oval. But then *so will an oval* when viewed from above. A *well-posed* problem would be like the problem an artist has rendering a 3 dimensional curves into 2 dimensions—this is a solvable problem, as there is a (mathematical) function to project 3 dimensions onto 2. Given a circular tabletop, and a specific viewing angle, there is a single solution regarding how to represent it in two dimensions. However, the inverse problem is intractable: without knowledge of the viewing angle, in our tabletop case, we cannot simply reverse engineer a representation of an oval into a circular tabletop from a specific viewing angle—we actually are faced with infinitely many 3D curves that it might represent. The same problem will be faced in edge detection: information picked up by transducers[2] in the visual edge detection system representing depth or range will be lost in the transformation from 3 dimensions to 2. Similarly, there will be

non-unique solutions to luminance and color reflectivity: different colors under different lighting may well be represented with the same values by the visual system (e.g., hue, saturation, brightness)—but then how can those be reverse engineered to represent what's *really* under view? A good example of this is the "checker shadow" illusion, created by Adelson (2005; Fig. 1, below) where two patches that are actually identical shades can be perceived as "different" based on assumptions placed on the data by the visual system to account for shade effects.[3] In the figure below, the areas marked A and B are identical shades of gray—though it's impossible to *see* that without masking the checkerboard pattern surrounding them. The visual system imposes some assumptions about both shade and patterns in order to interpret the image, and gets it "wrong" as a result.

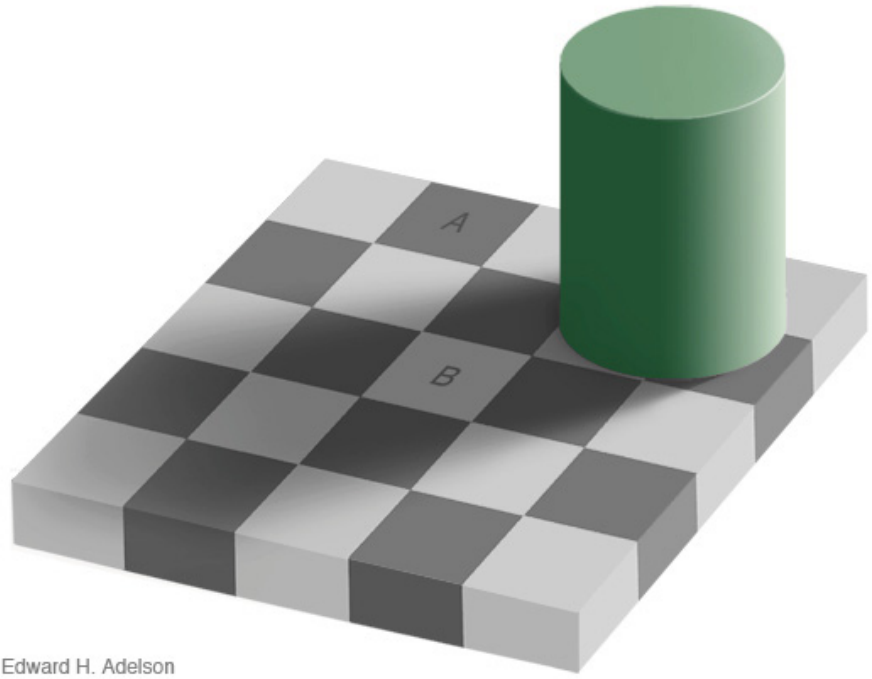


Fig. 1. The checker shadow illusion. Adelson (2005).

Marr (1982) notes that ill-posedness constitutes a potential framing problem at the heart of visual representation: namely, how is the system capable of outputting 'constancies' despite vast variation in proximal stimuli? It's bad enough with still images, given the non-uniqueness of the 'solution'—but even worse when we get to moving objects (or moving *subjects*): how does the visual system track objects in motion as maintaining shape when perspective is in constant flux? Marr specifically criticizes Gibson (1979) who claims that, somehow, "rigidity is specified"; Marr argues that the process cannot be anything like that

simple, but is nevertheless a matter of information processing within the visual apparatus (Marr 1982: 30). The solution, for Marr, Poggio, and many others, is to invoke some sort of "regularization" algorithms into the system—certain "assumptions" that are imposed on the data in order to essentially "fill-in" the gaps. These assumptions are innate parameters that have evolved with perceptual systems precisely because they constrain data in *useful* ways that allow fast and efficient and most of all *effective* representations of the external environment. One key feature of such regularization techniques is that, if they are to be effective, they need to be "encapsulated" within the system—which is to say, they cannot be interfered with by other systems. This is the heart of what it means to have a "modular" system. In the next section, I will explain what modularity entails, and bring it to bear on vision, as we have been discussing.

3.1 Modularity at the sensory periphery

The visual system as described by Marr (1982) operates as what is commonly called a *module.* As a first pass at definition, a "module" can be seen at the most basic level as a computational subroutine: an algorithmic mechanism tasked with a specific and clearly delineated processing task. A module takes inputs, runs them through according to its program, and spits out an output. The key element in that description is that information is processed by a module "according to its program". The "program" in the case of the visual module is what we mean by the *assumptions* or "regularization" placed on the incoming, inchoate data, as described above—the program of a module is designed to counteract the ill-posedness of the problem of visual perception. As Fodor (2000) explains, the frame problem of illposedness is "solved" by the module insofar 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).

Here is the basic idea: various subcomponents—transducers—of the eye convert (transduce) proximal stimuli from the world into computable representational format. For the visual system, this requires the detection and transduction of light frequencies and intensities (and for color vision: hue, saturation, brightness, contrast) in order to run *edge detection* and further image resolution algorithms on the data to isolate features (Marr, 1982). Image intensity values are transduced into mathematical quantities, and then computed in a

way that solves for *zero-crossings*, which are indicators of edges, or boundaries of the image (and presumably, of objects, or object features).[4] As noted above, this information is now in a syntactic form that is *computable,* but given the ill-posed nature of the problem of using it to represent the 3D world, that computability is compromised by the specter of combinatorial explosion. So the module has to have built-in limiting procedures—some way of imposing regularities on the data to maintain constancies *despite* small shifts or noise in the incoming stimuli. The data needs to be computable not just in principle, but in reality—respecting the 'finitary predicament'—which means its computability must be rendered tractable.

On Marr's view, (citing Poggio) extremely complex vision-guided behaviour *can* be modeled computationally with only a few variables and parameters, and without any 'top-down' processing, if one assumes a modular algorithmic architecture where the visual system processes variously transduced inputs in order to output simple, useful data for further integrative processing. Marr discusses the visual system of the simple housefly, and explains how its modular organization allows the fly a simple and effective way to process visual flight control, without much in the way of cognitive resources. Roughly speaking, the fly's visual apparatus controls its flight through a collection of about five independent, rigidly inflexible, very fast responding systems. For example, one of these systems is the landing system; if the visual field 'explodes' fast enough (because a surface looms nearby), the fly automatically 'lands' towards its center. If this center is above the fly, the fly inverts to land upside down. When the feet touch, power to the wings is cut off (Marr, 1982: 33). Of course, fly vision is obviously simpler than human vision, but Marr argues the same principles are at work, and the human visual apparatus likely "incorporates subsystems not unlike the fly's" (34). This leads to Marr to assume that our own visual system must operate via modules to allow for the "ease" of operation that we take for granted, despite the formidable informational complexity and processing demands of vision in a dynamic world. He argues for the modularity of human visual system based on the experimental isolability of various levels of function: "If we can experimentally isolate a process and show that it can still work well, then it cannot require complex interactions with other parts of vision" (Marr, 1982: 101).[5] This is the essence of what we refer to as the *encapsulation* of the system: its algorithm is not accessible to other systems, and not affected by other systems. The encapsulated module does what it is programmed to do, automatically, and with no

complicating influence or interference. This is how tractability of processing is maintained and frame problems are obviated.

So far, we have discussed just vision, but the idea is that all perceptual work must go on in a number of disparate dedicated perceptual modules, as opposed to in one single central, generalized 'world representation' device, as the latter would be intractable and slow, whereas survivability of cognizers requires fast and effective world-representation functionality across all sensory domains.[6] Fodor (1983) has made a strong argument that the entire perceptual system must be modular—encapsulated and highly domain-specific—if perception is to be computationally tractable and usefully represent the world. Of course, as is often the case, there is some terminological dispute as to what exactly it means for a system to be 'modular'—indeed, one of the central questions of this chapter involves the criteria by which one can count a functional cognitive mechanism as modular, and by extension how extensively modular the mind might be—so I should begin by clarifying exactly what I am taking the term 'modular' to mean, before posing the question of how extensively the human mind is modular. I will take Fodor's initial (1983) formulation as the starting point for this discussion.

3.1.1 Fodorian modules

For Fodor (1983), a module is a functional mechanism, but not just any functionally individuated mechanism; a Fodorian module is qualified by very specific criteria.[7] Fodor initially lists five questions that must be answered to determine whether a cognitive system can be considered modular (Fodor, 1983: 36-37):

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?

In answer, Fodor states that: "modular cognitive systems are domain specific, innately specified, hardwired, autonomous, and not assembled" (37). Unpacking these criteria a little further, Fodor's account is generally interpreted to list 9 attributes that jointly characterize modules. Modules are:

1. Mandatory in operation – activation is automatic and unconscious
2. Fast – processing is strictly limited and therefore happens quickly/reflexively
3. Domain specific – they operate over a restricted input domain
4. Informationally encapsulated – i.e., not "cognitively penetrable"
5. Inaccessible to central systems – (due to encapsulation)
6. Hardwired – they have a fixed neural architecture
7. Have shallow output – representational output is "simple", i.e., unelaborated
8. Develop according to a characteristic ontogenetic pace and sequence
9. Prone to characteristic and predictable breakdown patterns

Perhaps the key characteristics are the *domain-specificity, inaccessibility,* and *informational encapsulation* of modules. A module draws inputs only from a specified and restricted domain, and its processing is impenetrable to the rest of the mind; no information or algorithm from outside of the specified domain can be accessed by the modular processor or brought to bear on its operation. The input-output system essentially forms a computational 'black box' where feedback and exchange with other cognitive mechanisms is cut off.

> 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. (Fodor, 2000: 63-64)

Fodor holds up sensory input systems, such as vision, as clear examples of such encapsulated modular mechanisms. Visual input is quite clearly encapsulated, a fact that can be demonstrated by the characteristic patterns of breakdown visual perception is susceptible to—what we commonly refer to as optical illusions. Characteristic or predictable patterns of breakdown are often used as the empirical telltale of modular processing: the idea being that since the encapsulated processor is a black box, it will stubbornly (one might say stupidly) output the same response apparently oblivious to other information, belief or knowledge elsewhere in the cognitive system that the inputs are being manipulated or distorted in some

way. Once one understands the basics of how a particular encapsulated mechanism works, one can purposely misfeed it information and predict the response, which will be entirely 'wrong' in some broader general sense, but modules don't make such *judgments*—they simply follow the program, hell or high water.

The Müller-Lyer illusion presented in Figure 2 is a prime example:

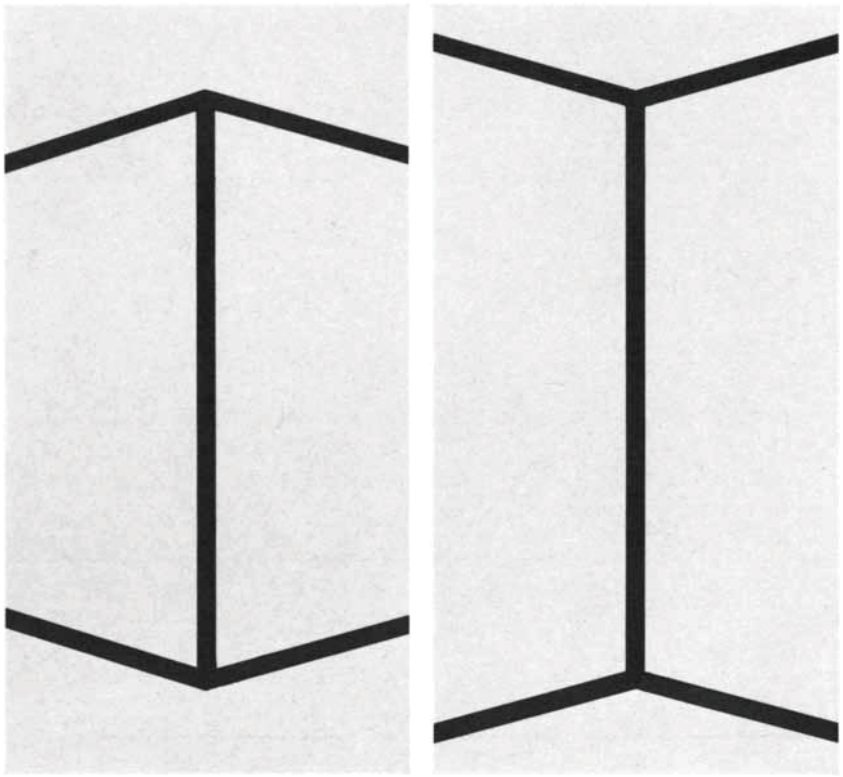


Fig. 2: The Müller-Lyer illusion. Taken from Gregory (1968: 70).

Despite the lines being the same length, our visual system is locked into outputting the right line as 'longer'. The arrows trigger an edge-detection function in our 3-D visual perceptual apparatus that represents the right line as *farther away* and generates an unconscious inference that it hence must be *longer*. The idea here is that the left line appears to the visual system as a convex edge, while the right line appears as a receding edge (Gregory, 1968). Figure 3, below, shows an example of a real world incidence of a Müller-Lyer style illusion:

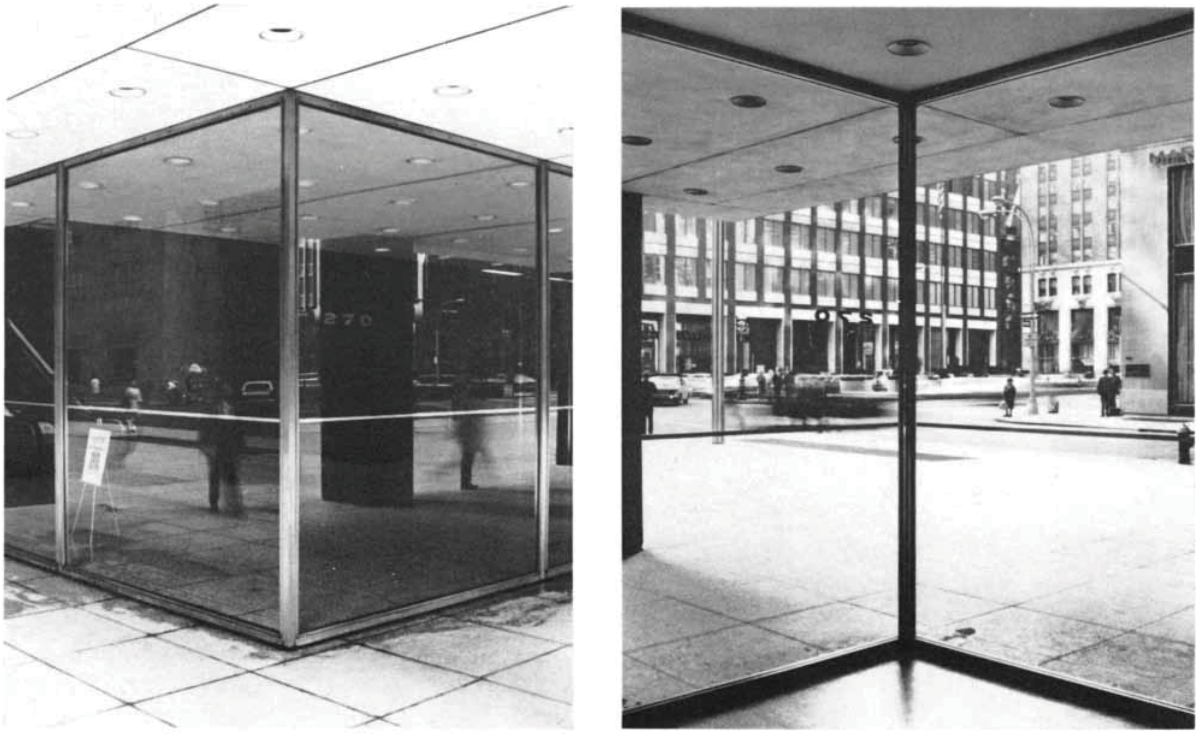


Fig. 3: A real world Müller-Lyer. Taken from Gregory (1968: 71).

Such visual cues are entirely useful in determining distance in the real world, and as a result, there is a perfectly plausible adaptive story for why we have evolved a visual system prone to this sort of illusion: the size/distance ratio of a perceived threat is an important piece of information to process, and any mechanism which speeds up that process (perhaps by employing some sort of Bayesian inference algorithm) will have immediate and dramatic adaptive payoffs.[8]

The interesting facet of the Müller-Lyer illusion with regard to encapsulation is the fact that the illusion persists even when it is known to be illusory. 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. Why can't the visual system 'recognize' that the Müller-Lyer image is only two dimensional—a drawing—and thus the 'normal' depth/size inferences should not apply? The clear answer is that the 2D figure is not anything that would appear in the natural environment in which our current visual system evolved—two-dimensional drawings are an extremely recent development, evolutionarily speaking. So the visual system has no program to account for the 2D/3D

dichotomy in this case, and treats all such images as if 3D. Indeed, as discussed above regarding the ill-posed and inverse nature of the problem of visual representation, the Müller-Lyer image is a perfect example of how the visual system is attempting to correct for the lost information—yet in this case *there is no lost information!* The Müller-Lyer image is a 2D image from the start—but the visual system "assumes" the 2D retinal images are meant to represent 3D objects, and hence it picks up on small cues in the data (in this case, the arrows) to project the 2D image back out into 3D. The fact that you "know" it's not 3D is not penetrating that module, so the data "correction" happens, over and over, unstoppably.

For Fodor, this proves that "perceptual processes are 'synchronically' impenetrable by—insensitive to—much of the perceiver's background knowledge" (Fodor, 1984: 39), which is to say that information from elsewhere in the perceiver's epistemic background cannot intrude on or affect the processing of the module as it happens in real time. 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. We can predict precisely under what circumstances it will return an "incorrect" answer, and we can construct various illusions to exploit this vulnerability.[9]

3.1.2 Why modularity?

Encapsulation—despite predisposing systems to some predictable breakdowns—has a clear an upside for cognitive processing. "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. It should go without saying that there is an evolutionary advantage to creatures that don't bog down computing sensory inputs, but are instead hardwired to interpret certain sensory cues reflexively in order to react quickly based on the 'information' provided. And *distance to percept / size of percept* would count as a fairly crucial piece of information when it comes to survivability in a hostile world. Crucial here is also the fact that what is output by the module is not a belief, but merely information which can then be refined by other, higher level systems and result in belief formation. In

this way, we *see* the Müller-Lyer lines as unequal in length, but we can *know* they are not, after bringing other cognitive resources to bear on the question. But this latter step takes time, and additional cognitive energy. If we take sensorial outputs at 'face value', we can react quickly *as if* the information the visual system outputs is veridical. And even once we have made the additional cognitive steps to form a belief that the lines are in fact equal, this new information is forever out of reach of the initial processing module. The lines will always *appear* unequal. You can't teach a module anything.

The alternative, of course, courts disaster. If I find myself with a hungry lion at my 2:00, and a potential safe-harbor cave at my 10:00, I am going to need to deduce (very quickly) a couple of key pieces of information: 1) How big is that lion? 2) How big is that cave opening? 3) Which is closer to me? It's probably better that I do not need to consciously work all that out, and am not forced to scroll through all potentially relevant information in my epistemic background. If I did do that, I would dither myself to death (i.e., like Hamlet). Instead, my visual system 'tells' me some basic things, quite automatically: 1) How big the lion is; 2) How big the cave opening is; 3) Which is closer. And I react. The information may well be factually incorrect in any particular instance. But if the encapsulated algorithm by which that information is generated produces statistically 'good enough' information to promote survival, then, from the perspective of parsimonious engineering, it is a much better design than a system that more often ends up with factually correct information, but requires vastly more (or infinite) time to compute possibilities.[10]

Fodor's (1983) contention is that "informational encapsulation is arguably a pervasive feature" of input systems (85)—systems whose dedicated purpose is to transduce and represent the "arrangement of *things in the world"* (42). Fodor stresses the computational obviousness of the natural wisdom of such an architectural solution:

> Proximal variation is very often misleading; the world is, in general, considerably more stable than are its projections onto the surface of transducers. Constancies [such as those output "stupidly" by modular subroutines] correct for this, so that in general percepts correspond to distal layouts *better than* proximal stimuli do. But, of course, the work of the constancies would be undone unless the central systems that run behaviour were required largely to ignore the representations that encode *un*corrected proximal information. The obvious architectural solution is to allow central systems to access information engendered by proximal stimulation *only after* it has been run through the input analyzers. Which is to say that the central processes should have free access only to the *outputs* of perceptual processors, interlevels of perceptual processing being correspondingly opaque to higher cognitive systems (60).

The case for modular systems, featuring informationally encapsulated processing, at the sensory periphery seems, to me, to be quite convincing. In the next section we will examine some empirical evidence that supports the view.

3.1.3 Empirical evidence of modularity

The idea that (at least) peripheral sensory systems are modular has been very well received in the literature, and evidence for such modularity abounds. Indeed, much of humankind's most ancient cognitive structures—those we have in common with our animal ancestors—appear to exhibit the key features of modularity. Let's examine a few ostensibly modular perceptual functions in order to further evaluate the strength of the claim.

We have already discussed the apparently modular design of the visual system at the earliest stages of edge detection. There is also a great deal of evidence that further stages of visual processing also exhibit encapsulation effects, such as illusory contour interpolation and amodal shape completion tasks. Consider a familiar Kanizsa figure, as in Figure 4:[11]

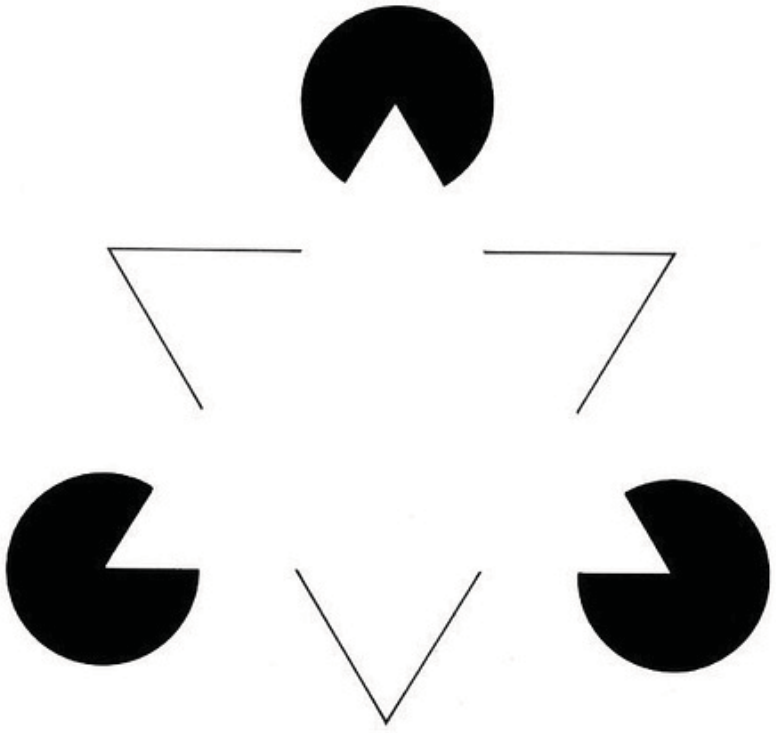


Fig. 4. A Kanizsa triangle. Taken from Kanizsa (1976: 156).

In the standard Kanizsa triangle, one cannot help but "see" the interpolated triangle, given the orientation of the 'Pacmen' and Vs. Additionally, the illusory central triangle is seen as *brighter* than the background. A similar effect is found in amodal shape completion (where a shape is occluded, and it is "filled-in" by the visual system).[12] Here is a classic example:

Fig. 5: The longest arm. Taken from Keane *et al.* (2012: 23).

The picture is comical because the visual system insists on treating player #13 as possessing one (very long) occluded arm. Knowing this *can't be the case* doesn't stop the interpolative process from completing the arm. Ringach & Shapely (1996) suggest that interpolative and shape completion functions "probably evolved to allow the observer to recognize and manipulate objects that are partially visible", given that "partial occlusion of objects is pervasive in everyday vision" (1996: 3048). They note that:

> In the process of object recognition, it is common for some boundary segments of the object to remain undetected. This could be due to low luminance contrast between the object and the background at some locations along the boundary, or because the object is occluded in some regions of the image. The visual system is therefore faced with the problem of linking the separate edge fragments that belong to the same object into a single unit. That a linking process takes place in the visual system may be observed

> directly in phenomena like illusory contours and amodal completion (completion of an occluded border behind an occlude). (Ringach & Shapely, 1996: 3037).

Keane *et al.* (2012) attacked the question directly whether contour interpolation is encapsulated, by checking to see if "*beliefs or expectations* can extinguish interpolation when it normally occurs or induce interpolation when it normally does not" (2012: 2). The verdict?

> Our result shows that filling-in contours happens automatically and that interpolation cannot easily be overruled by the beliefs that the observer has about the stimulus. In other words, not only does interpolation occur without thinking, it also occurs even when it is contrary to beliefs about contour connectedness... The outcomes of the current experiments add to the growing literature regarding the viability of the modularity research program. As Fodor (1983) rightly pointed out, it would be extremely helpful methodologically if the information processing systems of the mind turn out to be independent of central cognitive processing. It would mean that we could examine particular aspects of the mind, without regard to the goings-on of the beliefs and desires of the subject. If central characteristics of contour interpolation turn out to be intact irrespective of intentional states, then that would greatly simplify the quest to build an all-encompassing theory of visual object perception. Here, we have provided evidence that filling-in during interpolation, at least, does not strongly depend on belief. (2012: 14)

A further interesting finding of Keane *et al.,* is that there is a dissociation between *forming* vs. *noticing* contour interpolations, insofar as "94% of subjects exhibited behavior consistent with contour interpolation but only 62% reported seeing any such contours immediately after the experiment" (2012: 16). This suggests that the interpolation is automatic, and beneath the level of consciousness (as one should expect from a modular, encapsulated process), and that only when the attentional system draws the output from the interpolation function to the level of awareness is it phenomenologically reportable. Note that this is an important point which we will return to at length in later discussions about *integrative* modularity, and how the post-module outputs are processed at higher levels: for now, suffice it to say that the module does its work, but that work *may or may not* be taken up by further processing, at higher levels.

Another example of an ostensibly modular system which enjoys wide empirical support is that of linguistic parsing, both at the phonological and syntactic level. Fodor (1983) himself argues that there must be some sort of language "module", in order to pick up on wildly variable (incomplete, degraded) inputs, and sort out the signal from the noise—a phonological parsing system that can separate out the sounds of spoken language from all the background noise in order to process them effectively. Similarly, a syntactic processor has to

process linguistic items in such a way as to organize and prepare them for higher level (semantic) processing.[13] Fodor references Chomsky's (1959) refutation of Skinner and the "poverty of the stimulus" argument regarding language acquisition: clearly there are innate systems that impose syntactic rules on incoming signals that are not *learned*, as the stimulus is far too variable to explain how the robust language system that results could possibly be derived from the input patterns. Rather, the system must *impose* certain rules on the inchoate data, in order to usefully process it. Far too many studies in linguistics to cite suggest that syntactical and phonological parsing takes place via innate, mandatory, automatic and arguably encapsulated processes. Frazier (1987) argues that all the evidence suggests that syntactic analysis is "autonomous" and not connected to or influenced by semantic information or knowledge at initial stages. MacSwan (2013) agrees, and notes evidence that, in bilinguals, the syntactic parsing systems for the two languages show signs of being encapsulated even from one another.[14]

Similarly to linguistic parsing, lexical parsing in reading also exhibits signs of modular, encapsulated function, as evidenced by robust abilities to read text under severely degraded input conditions, though with interesting constraints. For example, people can generally very effortlessly read text in which the letters are jumbled, though only if the first and last letter are in the right place in each word; or subjects can read text in which letters have been replaced by similarly shaped numbers, though only after they have consciously recognized the switch. The example below is a classic: the task is simple, *count how many 'f's are present:*

FINISHED FILES ARE THE RE-
SULT OF YEARS OF SCIENTIFIC
STUDY COMBINED WITH
THE EXPERIENCE OF YEARS.

For the answer, check the footnote.[15] If you didn't count correctly, and many subjects will not count correctly, the standard explanation of your "failure" is that you ignored the Fs in the three instances of "OF"—why? Arguably because the "reading module" is skilled and automated to the extent that it can skip over connectives, and proceeds to do so, even when you are consciously trying to "read" specifically the Fs in the text. This is arguably the sort

of breakdown pattern (like in optical illusions) that suggests encapsulated function. Pelli & Tillman (2008) additionally have found that letter recognition is sensitive to "crowding" by other letters. Focus on the cross in the middle of this figure: when you do, the isolated **r** on the left should be clearly identifiable, while the **r** in "are" on the right will not be as clear, despite them being the same size, and same focal distance. Yet focus on the cross next to the "are" and now *both* **r**s are clear (Pelli & Tillman, 2008: 1130).

Fig. 6: The letter crowding effect. Recreated by the author.

Again, this suggests some sort of encapsulated processing *specifically* of letters. Note that *knowing* there is an **r** on the right doesn't make it any easier to *see* that **r**. This is very similar to the Müller-Lyer illusion in that respect

Yet another candidate for a modular reading is *music* perception. Peretz & Coltheart (2003) argue that evidence for a music module can be inferred from the experience of patients with *amusia*—the inability to hear melody, or to discern the difference between random and intentional strings of pitches. Amusic patients do not exhibit impairments in other areas of acoustic perception—they still understand speech, for example. In contrast, there are cases where trained musicians have lost the ability to speak or understand speech (due to brain damage) and yet retain their musical abilities. Such a double dissociation between the processing of melody and other auditory stimuli suggests a dedicated processor. Peretz and Coltheart suggest that the overarching music module is assembled from the parallel functioning of pitch organizing modules (contour analysis, interval analysis, and tonal encoding), temporal perception (which encodes rhythmic comprehension), and emotional and phonological modules which integrate the temporal and pitch information into 'meaningful' composites (since melodies are 'meaningful' as opposed to noise). The authors believe that at least some of these subcomponents can be proved to be modular, given the

existence of patients who are aphasic, yet can still sing, and congenital amusics who are incapable of parsing melodies, but can tap competently to rhythms.

Thus far, we have looked at peripheral sensory processes that are plausibly claimed to be modular. But what about more advanced forms of object recognition—those that would, at least on the surface, seem to involve coordination with higher level processes, belief, semantic knowledge etc.? There is some evidence that many innate aversions, such as those towards snakes for example, exist not only in humans, but are shared with other primates. Could *snake detection* be an encapsulated modular sub-function of the visual or object recognition system? Isbell (2006) argues that not only is snake detection an innate part of the visual system—insofar as even infants evince greater attention towards snakes and "snakelike" movements than other, less evolutionarily threatening creatures—but that "predation pressure from snakes has been a major force in the evolution of primate visual systems" (Isbell, 2006: 4). Literally, the primate visual system, according to Isbell, was designed in some respects specifically to deal with the task of quickly and effectively recognizing snakes. LoBue & DeLoache (2008), studying the speed with which infants can detect various stimuli, conclude that the evidence supports the positing of a "fear module":

> a bias in the detection of evolutionarily relevant threat stimuli very early in life... young children, like adults, detect snakes more quickly than three different kinds of threat-irrelevant stimuli (flowers, frogs, and caterpillars). There was remarkable similarity in the pattern of responses of the preschool children and their parents. These developmental findings are consistent with Ohman's (1993; Ohman & Mineka, 2001) proposed fear module—a neural system that is selectively sensitive to evolutionarily relevant threat stimuli. (LoBue & DeLoache, 2008: 14)

We can usefully connect this proposal to the evidence of Rozin *et al.,* discussed in chapter 1, above, regarding aversions to chocolate in the shape of feces, or potable water labeled "not sodium cyanide", or food touched by a sanitized cockroach. Perhaps there is a suite of "fear modules" that have evolved to pick up on various existential threats our early, and even primate, ancestors would have faced—a "disgust" module, for example. Arguably the entire phenomena of the "laws of sympathetic magic" (e.g., the *contagion* bias) could be the result of modular, encapsulated perceptual object recognition systems which pick up on key features that pose possible danger, and provoke aversive responses—responses which are, importantly, *fairly immune to correction.* Again, better to avoid something safe than fail to avoid something harmful.

Note, at this point, I recognize that we are veering into territory pejoratively dismissed as telling "just-so" stories. My point here is not to argue, specifically, for a modular disgust mechanism, for example. My aim is simply to note that many of the effects of *belief perseverance* discussed in chapter 1 may, *possibly,* be artifacts of encapsulated perceptual processes—*threat detectors*—that consistently output "danger", even when there is (more than) sufficient evidence to refute the "danger", and even when one is consciously aware of that evidence, and of the "irrationality" of one's response.[16] Of course, getting back to snakes, it is surely not irrational to fear snakes (indeed, it's arguably quite rational)—though there is a tendency to continue to show aversions to snakes (and spiders, and creepy-crawlies in general), even in situations where safety has been assured. Additionally, there is much evidence that the aversion to snakes, for example, is cross-culturally universal. Meanwhile, there is *no* innate aversion to automobiles—small children living in non-rural areas need to be *repeatedly* warned of the danger posed by cars in the street, whereas they do not need any warning, it seems, to back away from snakes. Yet, surely automobiles are a much more salient threat to the average city-dweller than snakes (which one might plausibly never encounter in one's entire life, living in the city). The simplest explanation for this would be that our perceptual systems, including post-perceptual object recognition systems, evolved in a particular environment—one where snakes were a worry, and cars not invented—and fast, efficient, *encapsulated* processes were selected for because they increased survivability. Keeping such processes encapsulated is the key to their success: no time to second-guess them.[17]

More generalized human abilities of object and facial recognition are also plausibly modular, as they have been shown to be very localized in the brain, and given certain types of damage, predictable behavioral patterns and cognitive defects present themselves. One example is the syndrome of prosopagnosia, or *faceblindness,* resulting from damage to the fusiform gyrus (the 'face area' of the brain).[18] Faceblind subjects recognize faces *as* faces, just not as any *particular* face. Frith (2007) points to prosopagnosia as the sign that facial processing is handled by a dedicated system, with all the hallmarks of being modular. Numerous experiments demonstrate that our ability to recognize faces works even under severely degraded input conditions—images can be distorted in numerous ways and yet be readily identifiable (Sinha *et al.* 2005). Non-face objects, or geographical features, are not

nearly as recognizable when the image is distorted. The inverted face illusion is yet another sign of this: when facial images are distorted, we tend to not notice the distortion unless the face is viewed upright (this is often referred to as the 'Margaret Thatcher illusion', as her face was the one used in initial tests). Our facial recognition system is not only input-limited to faces, but to *upright* faces, which makes some evolutionary sense—we don't generally find ourselves in pressing need of identifying upside-down faces.

More generalized object recognition is also arguably modular. Dickinson (1999) lays out an account of object recognition involving feature-grouping algorithms and some sort of 'hypothesis' generator as to what object best fits the data.

> The recognition algorithm evaluates, or verifies, each of the candidates in terms of how well it accounts for the image data. A score is typically assigned to each candidate, and the best-scoring candidate or hypothesis is chosen as the interpretation (or label) of the object. (Dickinson, 1999: 176)

Frith (2007) additionally highlights the neurological condition of agnosia, or 'object blindness'—literally 'loss of knowledge'—which results from damage to the occipitotemporal border and manifests in the patient's loss of ability to recognize previously known composite objects, despite retaining some ability to recognize constituent parts. Here, again, it seems that the mind has a dedicated module tasked with 'object composition', where the outputs of various sensory systems are grouped into objects from the bottom up, so to speak, first the 'parts' are processed (and identified, likely via some prototype-matching system), and then another system composes the conceptual parts into concept-wholes that are identified as such.[19] If this secondary system is damaged, the mind is left only with constellations of parts with no understanding as to what sort of object is constituted by them—e.g., a broom may be described as an 'oblong stick with hair on the end'.

Later, in chapter 5, we will discuss at length the proposal that many, or even *most* human belief, desire, thought and behaviour is subserved by encapsulated, modular processes, and in that discussion, I will highlight the so-called "massive modularity" thesis of Peter Carruthers (2006a) and others. For the time being, I want to focus on just a few aspects of Carruthers' account, specifically with regard to seemingly encapsulated *behavioral* functions of organisms, specifically insects. There are numerous examples of insect behaviour that follow what Carruthers calls "fixed action schemata"—behaviour that is

triggered by specific stimuli, and then proceeds according to a fixed response, regardless of whether the context is appropriate.

Carruthers highlights examples suggesting that much insect life operates on fixed-action schemata, which seem to resemble *de facto* encapsulated cognitive modules. One example of seemingly encapsulated *behaviour* patterns is that of the sphex, or "digger" wasp, as discussed by Dennett (1984): when the wasp returns to its burrow after a successful hunt, it leaves its prey at the edge of the nest, and proceeds to check inside whether another creature has occupied it, or is lying in wait. On the surface, this seems like fairly sensible animal reasoning—check for intruders before bringing in the catch—but the surprising finding is that if experimenters drag the prey a short distance away from the edge of the nest while the wasp is inside, when the wasp re-emerges, it drags the catch back to the edge and repeats the process (and can be made to repeat this loop *ad nauseum*). The upshot is that the wasp is getting stuck in a fixed action schema—there is a modular, encapsulated mechanism triggered by the environmental cues, which appears intelligent on the surface, but which can be manipulated to reveal itself as entirely programmatic and unintelligent, following a predictable recursive pattern.

Examples of modular insect cognition are also evident in the work of Gallistel (1990) on the 'dead reckoning' ability of ants. Ants can forage in a wildly circuitous fashion, and upon finding what they desire, return home in a straight line. If one experimentally manipulates the environment by letting the ant wander into a sandbox and then move the box some distance away while the ant forages, it will return on the exact vector that *would* return it home *if* the box had not been moved, and the ant is confused when home is not where it is supposed to be. No amount of auxiliary landmark information can sway the ant to re-evaluate the direction home: the dead reckoning ability is clearly encapsulated.[20]

Some degree of modularity is also plausibly suspected in mammalian spatial reorientation. Cheng & Gallistel (2005) have demonstrated that when rats are put in a rectilinear room and trained to find food in a certain corner, then removed from the room, disoriented, and placed back in the room (or one shaped just like it), they will consistently search for food in the 'same' corner or its mirror opposite. This might not seem so special, if it were not for the fact that when other identifying cues are given as to where the food is (one wall being blue, perhaps), the rats will still search in one of the two mirror image corners,

regardless of which one is accompanied by the salient landmark. The conclusion in this case is that reorientation works from geometric data in a modular fashion, and 'tells' the rat which corner to go to based solely on that information, impenetrable to other information regarding salient landmarks. The systematic breakdown here occurs because there is a mirror image corner to search—in the real world where this module evolved, few geometric features are truly rectilinear, so this confusion would seldom arise. Shusterman & Spelke (2005) have shown that this effect also works similarly with small children as it does in rats, but with the development of language comes an increased ability to override the geometric reorientation information in a diachronic fashion: the module still offers both corners as equally possible candidates to expect food (or toys in the case of children), but language capacity seems to offer children a chance to refine the information output from the reorientation module to rule out one of the two corners based on *other* salient information—something the rat is apparently incapable of.

3.2 Arguments and replies

The modularity thesis as applied to perceptual systems enjoys a fair bit of popular support in cognitive science, and many recent successes in computing—in computer vision, face recognition, speech recognition, text recognition, translation, etc.—are arguably attributable to the modular assumption, as such systems were designed to impose encapsulated regularization constraints, or "assumptions" on the data to make it computationally tractable, much in the way proposed for actual human systems (of vision, parsing, face recognition, etc.) However, there are numerous critiques of the modularity thesis, a few of which I will highlight in this section, though I argue that they are generally misplaced concerns. First, we will look at some arguments by Prinz (2006) that Fodor's focus on the importance of modularity has been overstated, and that there is evidence that *none* of the defining features of modular systems are (consistently) found in perceptual systems, and that where we do find them, they are uninteresting in terms of larger theory. In 3.2.2, I will look at broader arguments that numerous instances of so-called "cognitive penetration" violate the purported encapsulation of modular systems—and since, at least according to Fodor and Pylyshyn, encapsulation is the very heart of the appeal to modularity, this cognitive penetrability thereby renders the modularity thesis moot. I will argue in 3.2.3 that cognitive "penetration"

is actually no such thing, but, rather, evidence of higher-level integration and interface modularity. In other words, rather than being evidence *against* the modular thesis, I will argue that cases of so-called cognitive penetration are actually evidence of *further levels* of modular architecture, beyond the periphery. This will set us up for the discussion in chapters 4 and 5 regarding how far—or how *massively*—we can sensibly construe the modularity of mind.

3.2.1 Prinz's critique of modularity in general

Prinz (2006) argues that:

> When we draw boundaries around subsystems that satisfy any one of Fodor's criteria for modularity, we find, at best, scattered islands of modularity. If modules exist, they are few and far between. The kinds of systems that have been labeled modular by defenders of both peripheral and massive modularity probably don't qualify. Thus, modularity is not a very useful construct in doing mental cartography. (Prinz, 2006: 1)

In order to defend this claim, Prinz looks to the criteria postulated by Fodor (and Pylyshyn) for attribution of "modular" status, and argues that there are numerous ways in which even the systems presumed to be paradigm cases of modularity actually fail to meet some or many of the criteria. One example is the criterion of *localizability of structure*, which was #4 on Fodor's list: "Is it hardwired (in the sense of being associated with specific, localized, and elaborately structured neural systems)?" (Fodor, 1983: 55). Prinz argues that the evidence from lesion studies of localization of processing in the brain is much less clear than modularists make it out to be. He notes that "sometimes, lesions in the same area have different effects in different people, and all too often neuropsychologists draw general conclusions from individual case studies. This assumes localization rather than providing evidence for it" (2006: 3). Prinz recognizes that many systems in the brain do seem to have clear localization, but even when we *seem* to have located a specific function, we may still be overdrawing conclusions from mere localizability: as "when a lesion leads to an impairment of a capacity, we do not know if the locus of the lesion is the neural correlate of the capacity or the correlate of some ancillary prerequisite for the capacity" (*ibid).*

Similarly, when it comes to the question of *domain specificity* of operation (#1 on Fodor's list, above), Prinz argues that the empirical data is much more clouded than Fodor suggests. He flags the visual system as an example where *domain specificity* may often be

clearly violated.

> Consider vision. Edge detectors may be domain specific, but other resources used for processing visual information may be more general. For example, the visual system can be recruited in problem solving, as when one uses imagery to estimate where a carton of milk can squeeze into a crammed refrigerator. Some of our conceptual knowledge may be stored in the form of visual records. We know that damage to visual areas can disrupt conceptual competence (Martin & Chao, 2001). I have also noted that, when people lose their sense of sight, areas once used for vision get used for touch. Visually perceived stimuli also generate activity in cells that are bimodal. The very same cells are used by the touch system and the auditory system. If we excluded rules and representations that can be used for something other than deriving information from light, the boundaries of the "visual system" would shrink considerably. At the neural level of description, it is possible that only isolated islands of cells would remain. This would be a strange way to carve up the mind.... Vision, taken as a coherent whole, is not domain specific in the strong sense, even if it contains some rules and representations that are. (Prinz, 2006: 7-8)

I am not sure I see why the examples Prinz supplies here actually violate the purported domain-specificity of the visual system—to say that visual systems are "recruited" for higher level processing seems confusing. Surely, the *outputs* of visual systems are used in higher-level processing, such as in estimating what one can fit in the fridge. And arguably, there may be ongoing *re-activation* of the visual system to provide additional information (i.e., attention can direct one's eyes back to the fridge to "look for" additional information to aid in processing *what to do)*. But this doesn't imply the visual system, itself, is accessing other content domains. Rather, it merely suggests that, given other states of the organism, the visual system can be directed (and redirected) to provide information that can be used elsewhere up the chain. The domain of vision is still the proximal stimuli picked up and projected onto the retina, and the visual system, as a whole, stays firmly within that domain, regardless of how it is "recruited". As for visual images being stored, again, that seems a post-output stage, and hence doesn't bear on the question of the domain-specificity of the module *pre*-output. Similarly, the point that damage to visual processing affects cognition points merely to post-output processing: the visual system gives what it gives—if it is damaged, then it ceases to give, or gives corrupted data up the chain, which will show up in corrupted cognitive processing at higher levels. Indeed, as I will argue later in this dissertation, *this is exactly* why errors of higher order thinking (irrational thinking, belief revision failures, biased reasoning, etc.) could be the inevitable (and often incorrigible) result

of distorted or corrupted modular processing at lower levels: essentially, *garbage in, garbage out.* (Or: distortion in, distortion out.)

Prinz's final objections focus on the encapsulation criterion, which he suggests is routinely violated in examples where "top-down effects" can change, distort, or manufacture perceptual experience.

> For example, expectations can lead us to experience things that aren't there. If you are waiting for a visitor, every little sound may be mistaken for a knock on the door. Or consider visual search: when looking for a Kodak film carton, small yellow objects pop out in that visual field. The most obvious case of top-down influence is mental imagery. Cognitive states can be used to actively construct perceptual representations. This makes sense of the neuroanatomy: there are dense neural pathways from centers of higher brain function into perception centers. (Prinz, 2006: 9)

These sort of "top-down effects" are cited by many as fundamental flaws in the modularity thesis, as they seem to directly refute the encapsulation of modules. I will turn to these examples in the next section, including some others from Prinz.

3.2.2 Cognitive penetrability

On the surface, instances where "top-down" processes seem to influence the processing of perceptual systems seem to be a clear violation of the premise that modules are encapsulated—which, as discussed above, means both that the processing algorithms of the module are inaccessible (i.e., not introspectible) to higher-level systems, and impenetrable by other systems (i.e., they cannot be overridden). This was considered to be arguably the key virtue of modularity, as it explains how computational tractability can be maintained, free of outside "influence". But if the "black box" isn't quite so opaque, and is routinely cracked open—*cognitively penetrated*—and its processing is changed or blocked or affected by other cognitive systems, then that system shouldn't be deemed encapsulated, and hence the argument for its modularity vanishes.[21] Below is a quick hit list of various phenomena in which it is (or might plausibly be) claimed that an allegedly modular system is penetrated by another cognitive system. In each case, I have noted the purported "violation" of encapsulation. However, please note that this is not my own interpretation of these phenomena, and I will argue later that these are *not*, in the end, violations of encapsulation.

- **The McGurk Effect** (McGurk & McDonald, 1976)—when viewing a video of a person visibly saying /ga/ dubbed with audio of the person saying /ba/, subjects will report hearing /da/. As soon as they close their eyes, they hear /ba/. The illusion returns as soon as they look again. *Violation: parsing is penetrated by visual data.*[22]
- **Phoneme restoration** (Warren, 1970; Warren & Warren, 1970; Elman & McClelland, 1988; Prinz 2006)—when people are listening to sentences with missing phonemes (e.g., "the __eel fell off the bus"), they will "hear" the missing phoneme, as it is "filled in" (restored) based on semantic knowledge and lexical understanding of the sentence. However, a missing phoneme in a string of nonsensical gibberish is *not* similarly restored. *Violation: phonological parsing is penetrated by semantic knowledge.*
- **Orofacial manipulations on hearing** (Ito *et al.* 2009)—when listening to an ambiguous vowel sound (between a longer or shorter /a/), perception will be affected by stretching one's face: if pulled back, as when saying a longer /a/ as in "head", that is what will be heard. When pushed forward as when saying short /a/ as in "had", it will be heard accordingly.[23] *Violation: parsing is penetrated by orofacial proprioception.*
- **The red hearts experiment** (Delk & Fillenbaum, 1964; MacPherson, 2012)—judgments of color can vary based on what shape it is. People will generally judge heart shapes as *redder* than non-hearts actually the same color. *Violation: visual perception is penetrated by semantic knowledge*
- **The crossing/bouncing ball effect** (Sekuler *et al.,* 1997; Metzger, 1934)—a 2D representation of two dots crossing paths in an X pattern will be seen as bouncing off of one another if accompanied by an impact sound, but seen as crossing through one another with no sound. *Violation: visual perception is penetrated by auditory perception.*
- **The ventriloquism illusion** (Vroomen & de Gelder, 2004)—when an audio and visual stimulus are synched, but spatially dislocated from one another, the sound will be perceived as emanating from the location of the visual stimulus. *Violation: auditory perception is penetrated by visual perception.*
- **Touch-sound illusion** (Hötting & Röder, 2004; Prinz, 2006)—single taps to the arm will be felt as multiple taps if accompanied by multiple tones. *Violation: haptic perception is penetrated by auditory perception.*

Prinz certainly takes such effects to be violations of encapsulation, suggesting that "these examples show that there can be direct and content-specific cross-talk between the senses. The empirical evidence suggests that mental systems are not encapsulated" (2006: 11-12).

I do not think any of these cases point to violations of encapsulation. Rather, I believe they serve as evidence of *further modular systems*—systems that fuse or integrate representations from lower-level systems in order to create more rich sensory images (as well

as running error correction, filtering out "bad" or noisy outputs from sensory systems by cross-referencing them with other systems, based on expected synchronizations). In short, what looks like cognitive penetration of sensory systems is actually just the *predictable output* of dyadic modular cross-modal integration devices[24]*—themselves encapsulated.*

My argument for this turns on a relatively simple point: without positing modularity of integration, rather than cognitive penetration, one has no way to explain the fact that the ostensible "penetrations" are *uniform*. If it were cognitive penetration at work, one would expect it to be override-apt, for one, and for there to be differences in the "output" of the penetrated system, rather than uniformity across subjects and instances. In simpler terms, we should expect that if cognitive penetration is the issue, then perceivers should be able to cognitively penetrate the perceptual modules *differently,* or even *not at all.* In the purported instances of cognitive penetration, we are to presume that higher level systems, including semantic knowledge and belief systems, are *infiltrating* perceptual systems and affecting the output of those systems (in a way that violates modularity by definition). As Prinz states it above, there is "content-specific cross-talk" between various systems. However, the problem with viewing these phenomena as the result of "cross-talk" or as the effect of interacting with semantic knowledge and belief, is that such penetrations of the perceptual system in question should lead to *variable* and *unpredictable* outputs. The introduction of more global accessibility relations between systems vastly increases the complexity of the interactions, which should increase the variability of responses. The very *last* thing we should expect if there is "cross-talk" between systems, and infiltration of one system by another, is that the resulting output is still *uniform* across subjects.

Take the McGurk effect as a clear example—one that I will return to below in much more detail—the cognitive penetration account of the McGurk effect has it that somehow the visual system is penetrating and overriding the auditory system. However, it's not that simple, as we have seen: the "answer" that the visual system imposes on the auditory system is *wrong* (it comes up with /da/ rather than the /ga/ that is actually presented to vision). The question is: *why the mistake?* And worse, *why is the mistake uniform?* It would be one thing if everyone's visual system overrode their auditory system in this case and installed the *correct* response—at least that is a uniformity of response that makes sense, as there is only one right answer, after all. But a uniformly *wrong* answer is the last thing we should expect

from a process in which one system penetrates and overrides another: because there are potentially *infinitely many wrong answers.* Cross-talk between systems should result in highly variable responses, especially, at the very least, highly variable *mistakes*. But this is not what we find. On the other hand, if we substitute integrative modules for that cross-talk, we should expect exactly what happens: highly predictable patterns of response, and specifically, highly predictable and systematic patterns of *breakdown.*

In addition to not being able to account for the uniformity of response, the cognitive penetration account also fails to predict, and cannot account for, the fact that the effects listed above cannot be *overridden*. E.g., knowing that the video is of a man articulating /ga/ doesn't let your visual system "tell" your ears anything different; similarly, knowing it's a ventriloquist doesn't *reduce* the resulting impression that the sound is coming from the dummy; being shown that the heart really is the exact same colour as the background doesn't let your semantic knowledge *stop* telling your eyes the heart is redder, etc. If cognitive penetration, especially of background knowledge into perceptual systems, is the culprit, we should expect to be able to *override* it: to elect *not* to let our knowledge influence the perception, or to change the way in which our knowledge affects the perception. But in all the cases listed above, the effects are robust: "knowing" the illusion does not reduce it. So the cognitive penetration account makes the wrong prediction. On the other hand, the supposition that there is a modular, encapsulated, post-perceptual integration system, predicts exactly what we find: illusory effects that *cannot* be overridden, despite knowledge. The only way to override the illusion, is to mask one of the underlying perceptual inputs to the integrator. Hence, my contention that what *appears* to be cognitive penetration is actually evidence for post-perceptual modular integration: the modular account correctly predicts the results in all the cases of purported cognitive penetration, whereas the penetration account predicts a non-uniformity and override-aptness that we do not, in fact, find. In the next section, I will elaborate on this view building on arguments of Carruthers (2006a), Jackendoff (2002), and Pylyshyn (1999) to explain, or explain away, these cases of purported cognitive penetration.

3.2.3 Explaining (away) cognitive penetration

As Burnston & Cohen (2014) note, one way that many defenders of modularity might respond to seeming evidence of cognitive penetration is to weaken the standard of encapsulation required to count as a "module":[25]

> [O]stensible friends of modularity (in particular, defenders of massive modularity or evolutionary psychology more broadly speaking) have often taken evidence of the kinds of integration ... as reason for weakening or rejecting encapsulation as a criterion of modularity. Effectively, these theorists accept with the anti-modularists the idea that the evidence of informational integration refutes modularity *qua* classically conceived, and go on to "save" modularity by replacing it with something weaker. Thus, for example, Coltheart (1999) abandons encapsulation in favor of a loose notion of domain specificity. Sperber (2005) hopes to save modularity by distinguishing senses of domain-specificity relevant for understanding function from those that might be affected by interaction. Carruthers (2006) makes a similar move in offering a "weak" sense of encapsulation that can maintain functional specificity despite interaction effects from other processes. And Barrett and Kurzban (2006) distinguish between the information a module has "access" to and the information in processes. (Burnston & Cohen, 2014: 7)

In chapter 4 we will look specifically at some of these so-called "massively" modular accounts cited by Burston & Cohen. For the moment, I will appeal to Carruthers' (2006a) account of "wide-scope encapsulation", which he employs to explain the purported cases of cognitive penetration. Carruthers argues that "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, 2006a: 390ff). I am going to argue for a view very similar to this below, in §3.3 and at more length in chapter 4. Note that I think Burnston & Cohen's reading of Carruthers is somewhat off: Carruthers' goal is *not* to "weaken" encapsulation—the interaction systems he envisages *are* encapsulated—perhaps referring to them as "interactive" is problematic,[26] as it implies shiftable accessibility relations between subcomponents. My understanding of Carruthers' view is that what he is positing is very similar to the sort of *integration* systems I will defend below. Integration is a sort of "interaction" to be sure: but certain forms of interaction can be construed as entirely encapsulated—"black box" integration functions, if you will—activated (passively, mandatorily) by only certain pre-processed perceptual data from a highly specified domain at the input stage, processing that data according to an inaccessible algorithm, and culminating in shallow output (i.e., the output is unelaborated and only available for further processing by a limited number of higher-level processes). Jackendoff (2002) also discusses how

"virtually" modular faculties could be brought about: on his 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—he refers to this as "structure-constrained modularity" (2002: 205). This idea here is that 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 via bottlenecks, and the actual computational needs of the over-arching 'module' are made even simpler, and hence highly domain-specific.

As a potential example, take *snake detection*, as discussed above. In §3.1.3, I suggested that there is evidence that humans have a snake detection and aversion system that, at least on the surface, appears to operate as a module should be expected to (it is fast, apparently hardwired, automatic, seemingly impenetrable to other belief, prone to systematic "false positive" misfires, etc.). Such a system could easily be viewed as an *integrative* module operating on the outputs of two lower level modules: namely, a module that detects *snake-shapes* and a module that detects *motion*. Presumably, the invocation of a *motion detection* (modular) system in perception is likely not very controversial. Indeed, the existence of *akinetopsia*—a condition resulting from damage to area V5 in the visual cortex, in which patients are unable to perceive motion—suggests that there is a dedicated, isolable function responsible for processing motion (Frith, 2006: 26). The second lower-level system I am proposing is something that simply picks up on "snake-like shapes", long and roughly cylindrical. Now, in day-to-day perception, the motion detection system does not automatically trigger alarm and fear, and presumably neither would a "long cylindrical object" detection system (or else we'd be constantly recoiling at the sight of sticks and garden hoses). However, if we posit an integrative, bi-domain-specific *snake detection* module that is wired up to receive inputs from both (and only) motion and snake-shape detection, and activated *only* when simultaneously receiving inputs from said systems, then we can see how the bringing together of two modules via an integrator will end up processing *less* rather than *more.* This is the essence of what Jackendoff is talking about regarding how an input domain can widen in one sense, and end up narrowed as a result.[27] A *snake detection* system doesn't really do the work of motion detection AND snake-shape detection

and put them together. Rather, its domain would be *moving snake-shapes*, which is a relatively much more specified domain. Jackendoff concludes, referencing such interface and integration modules in language systems:

> 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 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).

Pylyshyn (1999) argues that the mentioned cases of seeming cognitive penetration are simply not what they seem, but rather the result of either a "within vision effect—i.e., visual interpretations computed by early vision affect other visual interpretations, separated either by space or time," (1999: 5) or a post-perceptual, higher level "judgment" stage, not a top-down infiltration into the lower level sensory modules as others claim:

> [I]f we view attention as being at least in part a post-perceptual process, so that it ranges over the outputs of the visual system, then there is room for much more complex forms of "perceptual learning", including learning to recognize paintings as genuine Rembrandts, learning to identify tumors in medical X-rays, and so on. But in that case the learning is not strictly in the visual system, but rather involves post-perceptual decision processes based on knowledge and experience, however tacit and unconscious these may be. (Pylyshyn, 1999: 33)

Pylyshyn, in his (1984) account, discusses how (essentially) modular systems can come online, *post*-perception, as assemblies of what he calls "compiled transducers"—where the outputs of one module feeds the inputs of the next in a linear process. Arguably, the visual system as described by Marr (and highlighted by Fodor as a paradigm example of an encapsulated system) is indeed a *compilation* of transducers, performing processes as various layers which feed one another.[28] On Pylyshyn's view, what *looks* like cognitive penetration of sensory modules, may in fact just be the effect of post-perceptual "compiled" systems, which have become computationally entrenched and automated to a degree that they function in a virtually encapsulated fashion—"indistinguishable" from prototypically encapsulated systems.

> [A] post-perceptual decision process can, with time and repetition, become automatized and cognitively impenetrable, and therefore indistinguishable from the encapsulated visual system. Such automatization creates what I have elsewhere (Pylyshyn, 1984) referred to as "compiled transducers". Compiling complex new transducers is a process by which post-perceptual processing can become part of perception. If the resulting process is cognitively impenetrable — and therefore systematically loses the ability to access long-term memory — then, according to the view being advocated in this paper, it becomes part of the visual system. Thus, according to the discontinuity theory, it is not unreasonable for complex processes to become part of the independent visual system over time. (Pylyshyn, 1999: 33)

I think this is the correct strategy for explaining (away) the purported cases of cognitive penetrability, in terms of those serving as the basis for objections to the modularity thesis. On the contrary, cases of cognitive penetration should serve to *support* the modularity thesis, and support the arguments for expanding that thesis beyond the periphery to include what I will provisionally refer to as *integrative modularity*. That is the argument I will pursue below, looking at each of the cases of ostensible cognitive penetration mentioned in the previous section in turn, beginning with the McGurk Effect.

3.3 Integrative modularity

I have stated above that I think what looks like cognitive penetration of perceptual modules is actually evidence of a second layer of modular system—namely, integration modules, tasked with resolving, composing, refining, and error-correcting the representations output by sensory modules. I will argue that these integration mechanisms will possess all the attributes of modularity—informational encapsulation, domain-specificity, fast, mandatory processing, etc.—the only difference between integration modules and lower-level sensory modules is the nature of the input. Whereas sensory modules are activated by physical properties in the world, picked up via transducers, integration modules take as inputs the representations output by various sensory modules. As I am envisioning, it is not that there is *one* integration system that take as input *all* the sensory information represented by perception—rather, I imagine that there are multiple systems (and levels, nested in hierarchies) of integration that take limited representational input from the senses, in order to clarify the perceptual 'scene' at any given moment.

The function of such integration modules would be primarily to construct representations across sensory domains, by fusing, cross-checking (for error correction), and

activating appropriate perceptual systems for further checking (or re-checking in the case of cross-modal mismatches). Let's look at the McGurk Effect as an example, as I propose it serves as evidence of an integration module between visual and auditory perceptual systems (and perhaps more, as I will explain).

Recall that in the standard McGurk illusion, the mismatched stimuli of lips articulating /ga/ while the audio recording is of /ba/ produces the perceived sound of /da/. As we have seen, some argue that this is some sort of proof of non-modularity, insofar as vision is somehow penetrating and overriding our perception of the phoneme actually recorded. As I mentioned in my brief outline of my argument against the cognitive penetration account in the previous section, I suggest that this is an implausible reading of the McGurk Effect, as it cannot adequately answer the following question: why does the override get it *wrong*, and *always in the same way?* In other words, if our eyes are simply overriding our ears, why are they mistaken (perceiving it as /da/ instead of what the visual image is *actually* articulating: /ga/)? And worse, for the penetration account, *why do others make exactly the same mistake?* Why isn't there some variability in the misfire—i.e., I hear /da/, you hear /la/, Timmy hears /ka/? The uniformity of the *mistaken* override demands explanation. I would argue that the uniformity of the mistake suggests a modular process—one shared with our conspecifics because it is hardwired. This certainly isn't a *cognitive* penetration, in the sense that a person *thinks* something like "those lips can't be saying /ba/ as there is no visible bilabial connection made, so it must be /da/" and subsequently interfering with the aural perception of the phoneme. Whatever is happening is happening very fast, beneath consciousness, automatically, immune to subsequent correction, and apparently in the same way across conspecifics. All of which suggests an evolved, hardwired processing algorithm dating back to common ancestors—in short, it's operating precisely as an encapsulated module should.[29]

Now, one might ask *why* such a module would have evolved—and with such a specific function: what use is there for a module that essentially splits the difference between sight and hearing—mixing audible /ba/ and visible /ga/ to create the experience of /da/? The answer: there is *no use for that.* The integrator would have evolved in an environment that did not include things like overdubbed videos. The *misfire* our integration module is exhibiting is precisely because it is being exposed to stimuli that are *unnatural* and

impossible in the actual environment it was optimized for. What the integration arguably *was* adapted to do is to crosscheck phonological parsing with some sort of visually mediated lip-reading in order to engage in noise correction and/or filling-in of gaps. In reality, the motor movements of /da/ and /ga/ actually *look* very similar: unless you are an expert lip-reader, you probably are not capable of resolving whether a certain pair of lips is articulating /da/ or /ga/ without further information. Similarly, as we can all attest, the *sound* of different phonemes can be easily mistaken, given only the acoustic cue. (Hence when spelling out things, or giving postal codes, we say things like "B, as in *Bob"*—to ensure it's not heard as something else, a P for instance—or consider the alphanumeric codes used by radio operators, *bravo, delta, golf, etc.)* So, finding ourselves in a world where lip-reading (alone) will often get stuck with ambiguous percepts, and phonological parsing (alone) will as well, it makes perfect sense that a cross-check, comparator system would be adaptive. Such a system would take as inputs precisely (and only) lip-reading and phonological parsing outputs. The integrator then compares the two, and, in cases of mismatch, imposes a "correction". It doesn't matter that the correction isn't perfect; for, it just needs to be good enough. In the *real* world, one might hypothesize that when the eyes see /ga/ and the ears hear /ba/, the integrator simply rules that it must *actually be* /da/.

Perhaps the integration system simply judges ¬/ba/ (and ¬/pa/), i.e., ¬bilabial, based on the visual input which *rules that out*. Then, of the choices that remain (i.e., those featuring alveolars **d** & **g**, and the velars **k** & **t**, all of which *look* pretty indistinguishable to a non-expert), the one that is *closest* according to some psychoacoustic parameter is selected and output. Essentially, all we have to assume is that the integration system a) prioritizes the visual cue, perhaps because it's statistically less prone to ambiguity error; b) using the visual cue, makes a "short list" of velars and alveolars; c) compares what the auditory system reported to the short list; and d) if there is no match, then the "closest" match to the auditory output that appears on the short list is duly assigned.[30] The system is impenetrable, so it can't be overridden, even *knowing* that the mismatch is untrue, or manipulated. And what we (consciously) "hear" is not the direct output of the auditory system: it is the product of further layers of post-sensory processing, refinement and error-correcting. So it's highly misleading to suggest that our *vision* changes the way we hear things. Rather, I would argue that *every*thing we "hear" has been crosschecked and integrated with other systems before we

consciously are aware of "hearing" it. Below is a simple diagram of the sort of workflow I am describing:

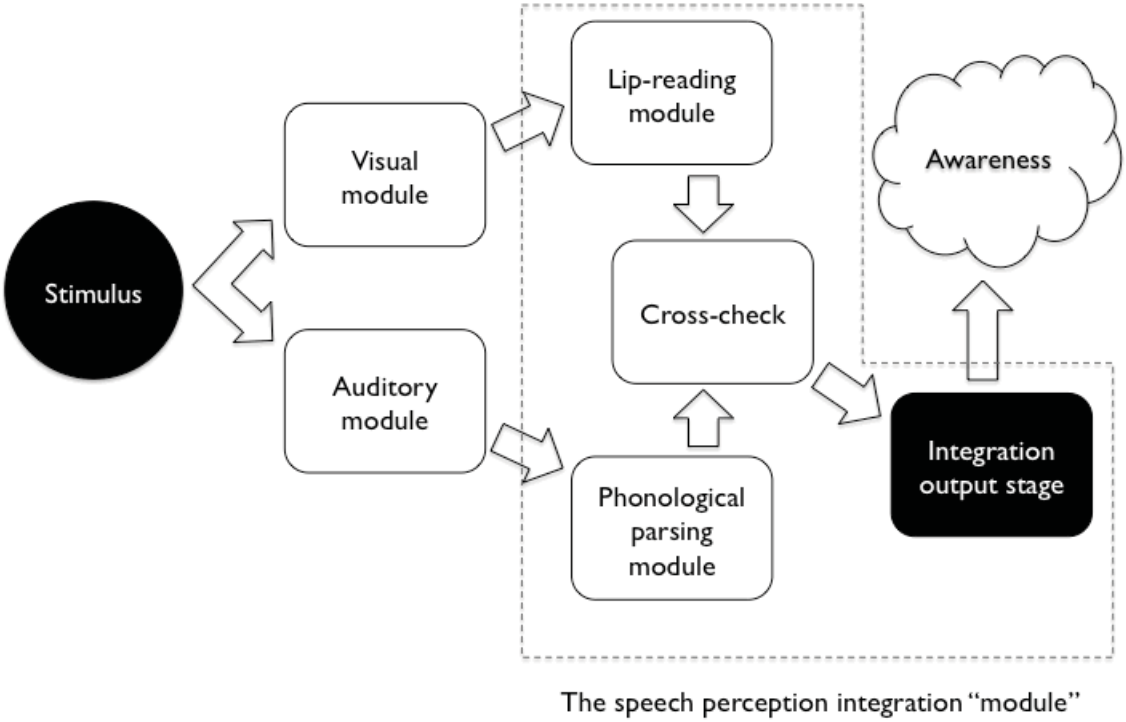


Fig. 7: A speech perception integration "module".

Note that in this diagram, the dotted line delineates what is essentially an encapsulated processing system: it takes as input only features relevant to speech perception (lip-reading and phonologically relevant acoustical cues), its internal processing cannot be accessed, and conscious awareness can only access the integrated output.[31] This would make a lot more sense, actually, than the idea that conscious awareness can *directly* access the raw data being output by the auditory system (or the visual or any other system): as we know, most of that information is ignored and discarded without further processing, let alone conscious awareness. I would suggest that conscious awareness can never *directly* access the representations output by perceptual systems—that's why those output are *shallow*—conscious awareness gets a sort of "executive report", later, after integration, error correction, and assorted representational fusing has all succeeded in framing a coherent 'scene'.[32]

The McGurk Effect is not the only "illusion" listed in §3.1 as evidence of cognitive penetration that could be better explained as evidence of modular integration processes. I won't give a thorough explanation for each, here, as I have above regarding the McGurk Effect. I will merely gesture quickly at the sort of integrator I would posit in each case, and suggest it is evolutionarily plausible:

Phoneme restoration – in this case, just as the visual system, including lip-reading, can help run error-correction on ambiguous phonological percepts, given a creature that has developed a language system, replete with semantic connection to the phonological cues, it seems highly plausible that a simple integrator, activated by ambiguous aural representations, could run those representations through a simple NOT-gate. E.g., in the case of "the __eel fell off the bus"—the language processor (remember: a prototypical *modular* system) could generate a "short list", just as in the McGurk Effect. From there, simple error correction can rule out what the missing phoneme *cannot* be, leaving only <wh>.[33] No actual violation of the parsing module is going on—just an encapsulated, automated *interpolation*. Note, if we already assume that *contour* interpolation is a modular function, as discussed above, then lexical interpolation seems just as likely—but it doesn't happen in the auditory module, it happens after, and consciousness can only access the *result*, not the earlier output stage. What we consciously perceive as having "heard" is not what our sensory systems outputs (it can't be: the <wh> was never there in the perceptual stimulus). Rather, what we hear is a *reconstruction* of what we essentially *must have heard*, after error correction, noise attenuation, interpolation and perceptual smoothing have "cleaned up" the data.[34]

Orofacial manipulations on hearing – this is the case where, when listening to an ambiguous vowel sound (between a long and short /a/), perception will be affected by stretching one's face: if pulled back, as when articulating a longer /a/ as in "head", that is what will be heard. When pushed forward as when articulating a shorter /a/ as in "had", it will be heard accordingly. Here, again, I would posit a simple error correction system designed to match lips with sounds. In this case, it's not a visual match, but a match based on proprioceptive feedback. The results of the Ito *et al.* study specifically suggested that the *motor theory of speech perception* could be vindicated by these results. (This is the theory that holds "the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements of the

articulators through certain linguistically significant configurations" (Liberman & Mattingly, 1985: 2). It also might simultaneously explain the McGurk result and the phoneme restoration effect: if every phonological perception triggers a sort of mental rehearsal of the motor required to *produce* that sound, then this could, in turn engage other systems, such as orofacial proprioception, semantic understanding, etc., that would all have been associated with speech production. Running all these systems in parallel, can allow for quick, automated filtering for noise and drive ambiguities to a discrete position, based on prior association.[35]

The red hearts experiment – color perception is *notoriously* ambiguous, prone to relativity effects, changes in lighting or surface reflectance, and the problem of accounting for the existence of metamers.[36] As a result, it's not surprising at all that color perception data would be picked up by other system to *impose additional assumptions on the data*—and then, under those assumptions, to run cross-checks for what *cannot* be. Assuming that heart shapes are usually associated with RED, this assumption can be imposed by semantic integration (it's *not* that it infiltrates the visual process, it's rather that another system *automatically* and *impenetrably* skews the percept towards the red. And note, it only does that in *comparison* to another patch or shape of color. I.e., *blue* hearts don't simply turn *red* in perception. Not even *purple* ones do. It's just that relative to other red (non-hearts), hearts look *redder.*

The crossing/bouncing ball effect – here we get what seems like a simple case of a perfectly ambiguous visual representation: like a Necker cube[37], or the spinning ballerina, we can only see it *as* oriented in one direction or the other. Whereas in the visual bimodal illusions like the ballerina, we can only pick between *visual* representations—so it's merely an either-or—in the bouncing balls case, we can run a crosscheck with *another* sensory system, and the confluence of sight and sound outputs *bounce* rather than *cross*. (Note that you *can "see"* the balls as bouncing, even without the sound. What is difficult is to see them as crossing *with* the sound: in the bimodal case, we get an encapsulation effect—a mandatory output.)

The ventriloquism illusion – when an audio and visual stimulus are synched, but spatially dislocated from one another, the sound will be perceived as emanating from the location of the visual stimulus. Here again, I suspect a simple error correction system: determining the

direction of a sound can be extremely difficult, whereas determining the direction of a visual stimulus is usually easy. So if the two are simultaneous, then an integration stage would quickly impose the directionality on the auditory percept. Again, this is *not* actually interfering with the auditory module and changing its output, but, rather, fusing the two perceptual outputs into a coherent whole using a *very* simple algorithm. And conscious awareness can access only *after* that point; for the effect is "baked in", after the initial perceptual modules, but before awareness.

Touch-sound illusion – where single taps to the arm will be felt as multiple taps if accompanied by multiple tones. I will offer only one speculative hypothesis: it's just a side effect of the same sort of system I have posited in the bouncing balls case above—some sort of automated impact and sound integration, which works pretty well in general, but gives this odd misfire (which has no seeming maladaptive cost!)

I am arguing that all of these cases of purported cognitive penetration are *better* explained as evidence of encapsulated integration systems. These effects exhibit little sign of cognition, yet all the characteristic misfire patterns and mandatory functioning of modular systems. Without positing modularity of integration, rather than cognitive penetration, one has no way to explain the fact that the ostensible "penetrations" are *uniform*. If it were cognitive penetration at work, one would expect it to be override-apt, for one, and for there to be differences in the "output" of the penetrated system, rather than uniformity across subjects and instances. One might interject here that some of the cases mentioned above *are* in fact "override-apt"—in the sense that, e.g. in the phoneme restoration case, one could re-listen to the prompt and hear the missing phoneme, once one had come to understand that it wasn't actually there. Or in the red hearts case, one might realize on a second look that it's no redder than the background. I would argue that those cases are actually different perceptually, given that attention is drawn to a specific detail, rather than the whole scene—as a result, it wouldn't be surprising that the scene is perceived differently, *when it is literally perceived differently.* What I mean when I say "not override-apt" is that when perceived in the effect-context, the effect will happen. Changing the attentional direction of the context *changes* the context—just as in the Müller-Lyer illusion, if you add a ruler in between the lines, you can suddenly see they are actually equal. That doesn't make the illusion "overridable", it makes it "work-aroundable". This idea of "working around" rather than

"overriding" will be a key point in my later arguments, especially in chapter 8 regarding delusional belief.

3.4. Barsalou's simulation theory

One account that describes how perception is the result of a multi-modal integration in a somewhat similar vein is that of Barsalou (1992; 2003; 2009). Barsalou's position is that perception is streamlined and sped up in its processing by the activation of memories of previous experiences with relevant feature associations/similarities. The memory will activate a "simulation" from which expectations can be generated, and subsequently compared to the incoming percept in order to smooth out irregularities, attenuate noise in the signal, rule out error, and interpolate where there are gaps. Barsalou explains how his account would explain phenomena such as the phoneme restoration effect:

> During auditory perception, lexical knowledge produces predictions via simulation that contribute to speech perception. In the phoneme restoration effect, listeners use auditory knowledge about a word's phonemes to simulate and predict a missing phoneme (e.g. Warren 1970). When a phoneme is missing, information present for surrounding phonemes is sufficient to activate a simulation of the word that includes the missing phoneme. According to the account proposed here, when a perceptual stimulus activates a similar perceptual memory, the perceptual memory runs as a simulation of the stimulus and speeds its processing by activating relevant processing areas, with the simulation perhaps fusing with the stimulus information... the memory predicts that the stimulus is another instance of itself, thereby increasing the fluency of perceiving it via top-down activation. (Barsalou, 1999: 1284-5)

Of course, "via top-down activation" seems to be precisely the explanation one is presumably hoping to *avoid* by positing modular integration systems. However, the key word here is *activation:* the "top-down" influence is not getting *into* the perceptual module in order to override it—the processing algorithm is not modulated in any way by the introduction of background knowledge (which would be a clear violation of informational encapsulation). Rather, the description is one where perceptual modules can merely be *activated* by top-down processes—essentially re-queried or selectively activated—instructed to "look again", so to speak. This does not violate the encapsulation of the module being re-activated, as there is no interference with the processing of the module, the top-down influence is indirect, by reactivating the lower level system to re-run the program. I would propose that first level, sensory modules can only be re-queried in a brute force way, by re-directing them to the

stimulus. So, for example, the visual system can be directed to *look again*, or hearing and parsing systems to *keep listening* to the rest of the stimuli (as maybe the *end* of the sentence will help unpack or clarify the *beginning* of the sentence). What the top-down system *can* do is selectively reactivate some systems, to indirectly affect the post-perceptual integration stages. As the sensory percept unrolls in real time, simulations can be run as to roughly *how this usually plays out*, according to past perceptual memory. The simulation can then help smooth over gaps with the incoming stimulus, and can direct attention, via redirecting senses, to focus on the details that would (as experience has taught the system) help resolve the signal.[38]

Beyond the periphery, if there are levels of sensory integration at work, as I have argued, then the selective activation of those modules could be done via simulation alone: if under normal perceptual conditions, visual representations are output by vision in a particular syntactic form (A) and representations of phonemes in another form (B)—which is just to say, they are interpretable as such—then there seems to be no reason why representations of *past* visual and parsing experiences could not also *feed the integration module inputs,* as presumably the memories could be coded in the same representational form as the original. In this sense, memories of past perceptual experiences can be layered on top of currently perceptions to resolve the image. (Note that this is much like how Marr (1982) describes the process of vision even within the visual system. All Barsalou is adding here is that memories activated via simulation can sketch the road ahead, and if the perceptual experience roughly fills in the sketched "road ahead", then small gaps in the information will be smoothed over.) This process would improve tractability immensely, insofar as small perturbations in the perceptual data could be largely *ignored* as long as the bigger (situational) picture was close enough to the simulation, thus freeing up cognitive resources. Barsalou notes that the automation that comes with expertise is largely a result of this sort of highly successful simulation-running: if the simulation is good, then the organism essentially *knows what is coming next*, and the motor responses can be activated ahead of time. Barsalou notes that:

> a situated conceptualization is a multi-modal simulation of a multi-component situation, with each modal component simulated in the respective neural system knowledge about these familiar situations becomes entrenched in memory, supporting skilled performance... Over time, the situated conceptualization becomes so well established that it becomes active automatically and immediately when the situation arises. (2009: 1284).

This explains certain findings such as neuroimaging studies that have shown that "when reading about a sport, experts produce motor activations that novices do not" (Barsalou, 2009: 1287). Similarly, research involving pattern recognition and memory in master-level chess players support Barsalou's view. 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 simulated, as they activate patterns previously processed and stored in memory, rather than positions that are not the result of actual play (and whose randomness fails to match with any simulations). In short, playing a lot of chess makes you not only better at playing chess, but at anticipating chess, and even at remembering chess-related perceptual stimuli.

We will return to Barsalou's account again, in the discussion of concept formation and how the concept acquisition, composition, and revision processes may be subserved by encapsulated processing in chapter 5. Barsalou's account is quite rich and has a number of further implications for how concepts are stored and retrieved using what he refers to as "conceptual frames" that will be helpful to the account I will propose later in this dissertation. For now, I cite him as a supporter of the idea that multi-modal integration is *part* of perceptual processing, but not via penetration of the sensory modules, but, rather, mediated by *largely automatic* integration systems. I have argued in this section that such integration systems are themselves modular—they are highly domain-specific in terms of inputs, they exhibit signs of informational encapsulation, given that they are prone to systematic patterns of breakdown, they are mandatory in operation, they are fast, and they are so prevalent and consistent in operation across the species that it makes sense to claim for them a particular ontogenetic status, especially given that one can tell a *very* plausible "just-so" story for their evolution.

3.5 Review and look ahead

In this chapter, we have looked at how the *frame problem*, at its most basic level of perceptual processing, is arguably solved by assuming that perceptual processes are largely

modular, informationally encapsulated and operate over highly specific domains. Modular systems that evolved in response to particular selective pressures (such as needing veridical representations of a dynamic, multi-dimensional environment across sensory modalities; threat detection; communication, etc.) are designed specifically to *impose* frugality on infinitely variable data, and output computationally tractable representations for higher-level processing. Informational encapsulation is a key element of this design, as inaccessibility of function preserves the speed of the system, even at the price of some "mistakes" under certain conditions—what we call "illusions" in many cases. Modularity establishes frames on inchoate perceptual data—based on the adaptive success of certain framing algorithms over evolutionary time periods—and thus the frame *problem* is not so much "solved" as merely obviated at the perceptual level.

We looked at some pushback to the this idea, including certain phenomena of "cognitive penetration" that seemingly violate the encapsulation of perceptual modules, but I have argued that what looks like cognitive penetration is actually more likely an indicator of *further* levels of modular processing—specifically, integrative modules that operate over the outputs of various modular subsystems, and are themselves, in turn, informationally encapsulated with respect to higher level systems.

In the next chapter, I will continue in this vein and argue that this account of sensory integration modules can be extended to include further levels of more broadly construed "assembled" modules, including, crucially, those involved in the sort of higher-level cognitive tasks that underwrite conceptual understanding, belief formation/revision, and inductive and abductive reasoning. This will include some discussion of what is known as the "massive modularity" thesis, which posits that modular structures are to be found underlying most, or even all, cognitive processes. Fodor, who we have turned to for the genesis and defense of the modularity thesis, famously decries the extension of that thesis to more "massive" levels. In the following chapter, we will address those concerns as well.

Notes for chapter 3

[1] As opposed to a "well-posed" problem, as initially defined by Hadamard (1923) as a problem whose solution exists, is unique, and is not sensitive to noisy data fluctuations (Poggio, 1985). Also see Poggio & Koch (1985) and Marr & Poggio (1979) for further discussion of "ill-posedness" and its relevance to early visual output. *Cf.* Kabanikhin (2008) for mathematical problems that illustrate illposedness.

[2] A word on transducers: I won't go into a lengthy discussion of transduction here, in the interest of brevity. But for a quick takeaway definition, Pylyshyn (1984) is helpful: "The typical transducer simply transforms or maps physical (spatiotemporal) events from one form to another in some consistent way" (Pylyshyn 1984: 151). The "other" form in the case of cognitive transducers means a representational form that can be taken up and processed by other cognitive mechanisms. Pylyshyn lists 3 criteria for sensory transducers: 1) they perform a "primitive, non-symbolic" function, and are simply a brute fact of cognitive architecture; 2) that it be stimulus-bound, and driven only by environmental triggers, independent of the rest of the cognitive system; and 3) that its behaviour is to be "described as a function from physical events to symbols" (1984: 153-4). See Dedrick (2009) for critical discussion of Pylyshyn's view.

[3] The problem goes both ways: on the other hand, we can have *metamers*, or instances of two patches of color than have quite different surface spectral reflectance distributions (SSRs), but are nevertheless perceived as "the same" by a given observer in a given viewing frame. See Cohen (2009) or Byrne & Hilbert (1987) for a full account of such phenomena, mostly looking at the metaphysical questions surrounding what this all says about color and color perception.

[4] I am not going into depth about Marr's levels of visual processing, as it takes us somewhat far afield of the central question of framing. See his (1982), or Poggio (1981) for a good summary of the view. In the quickest description, Marr argues that visual representation involves three levels of processing, in order to impose regularities and get around the ill-posedness issue. The first level involves quantifying image intensities and solving for zero-crossings to create the "primal sketch" which essentially isolates boundaries. After this, a "2.5D" image is constructed, which is basically an *egocentric* description of the image, which includes elements of surface contours, depth, etc. Finally, the third stage is a quasi-objective 3D "model" of the objects represented in the image, which can be manipulated and employed in cognition.

[5] *Cf.* Kowler (1999) for more on the modularity of vision including how algorithmic saccading contributes to a modular attention system.

[6] Also, the simple fact that senses are both functionally and neurologically dissociable shows that they process separately.

[7] 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 that will be addressed in the discussion of integrative or 'assembled' modularity, below.

[8] One might object here that Gregory's analysis of the Müller-Lyer seems odd, given that the sorts of edges that would have existed in adaptive environment of our ancestors wouldn't look much at all like these modern architectural examples. I think one might argue that modern architecture triggers the illusion rather easily, but that the depth/edge detection system was wired in to deal with much looser, less square, natural phenomena of the same sort: mainly cliff edges and the like. "How close am I to the edge of the ravine" etc. This is why it's so easy to get a false positive from that system – since architectural edges are simply the *perfect* versions of what we have evolved to spot in the wild, which actually helps explain why they are so impenetrable and robust as a distance illusion. And better to get a false positive from the system than the converse: better to fail-*safe* than fail-*dangerous*.

[9] Other examples abound: classic Ames room illusions, in which perspective effects distort height; or seemingly gravity-defying natural landscape illusions such as New Brunswick's "Magnetic Hill"; rotation illusions, such as the spinning ballerina, who can spin clockwise or counter-clockwise because the 2D animation is ambiguous; shade and color illusions, in which relative reflectances and backgrounds can make shades or colors appear different when they are not (or the same when they are different). See Michael Bach's webpage for a wealth of examples: <http://www.michaelbach.de/ot/>.

[10] As discussed above with reference to a 'satisficing' criterion, rather than maximizing. Always having the 'right' answer would be *ideal*, but if one ends up dead before the right answer comes, it's not much use. Better to have a 'close-enough' answer in time, than a perfect one too late. As I noted in the precious section, I would argue that *all* adapted mechanisms are designed in general to fail-*safe* in this way, rather than fail-*dangerous.*

[11] Cf. Kanizsa (1979; 1985) for a complete discussion of such figures.

[12] For a full discussion of the phenomena surrounding "filling-in" of visual perception, see Pessoa *et al.* (1998), which offers numerous examples.

[13] Note that Fodor doesn't explain much more about a language "module" than this, although we will expand on what exactly would be entailed by a language module later on—especially with regard to the question of whether it is a single module (unlikely, I will argue, citing Jackendoff, 1987) or a suite of interacting, yet separately encapsulated modules.

[14] Note that, below, we will examine some evidence that purports to show infiltration of "penetration" of the language system, though not necessarily at the syntactic level.

[15] There are 6 *F*s. If you only counted 3, you are not alone. The explanation is that conscious attention tends to skip over the connective "of", which appears three times.

[16] Again, from an engineering perspective, an alarm prone to err on the side of false positives rather than false negatives would be preferable if the goal is survival (e.g., reflexive ducking)—failsafe design. A threat detection system that operates in such a way that it sounds the alarm without ever waiting to check with other systems is more likely to survive that one that *does* take the time to verify the actuality or exact probability of the threat.

[17] This is often referred to as the "Pleistocene hypothesis" in evolutionary psychology: the idea that 99.9% of human evolution took place in a fairly static threat environment, so we evolved systems to deal with that environment. The modern world, being so recent, is not necessarily the world we are well-adapted for—different threats—and this fact reveals itself in the many cognitive biases to which were are prone. For much more on this topic, see Marcus (2009), Barkow, Cosmides & Tooby (1992), Richerson & Boyd (2005).

[18] See Frith (2007); Ramachandran (2011) for a full discussion of the neuroscientific evidence for this claim.

[19] A perfect example of the sort of modular "integrative" system that I will defend in chapters 4 and 5, below.

[20] Additional work suggesting modular cognitive structures in insect life can be found in Tarsitano and Jackson's (1997) research on the mental mapping abilities of with araneophagic jumping spiders, and Gould's & Gould's (1988; *cf.* Gould, 1990) work on honeybee dancing and social communication.

[21] I think there is some confusion in the literature as to how precisely to define cognitive penetrability. Siegel (2011) defines it as follows, with reference to vision:

> If visual experience is cognitively penetrable, then it is nomologically possible for two subjects (or for one subject in different counterfactual circumstances, or at different times) to have visual experiences with different contents while seeing *and attending to* the same distal stimuli under the same external conditions, as a result of differences in other cognitive (including affective) states. (Siegel, 2011: 205-206)

I think this is *not* a good definition at all, as there are already a lot of situations in which this is the case—e.g., color-blind people will have completely different visual experiences of color than I will, even under identical viewing conditions. That's not cognitive penetration, that's just perceptual difference. It can't just be invariant *external* conditions triggering distinct *internal* perceptual experiences that count as cognitive penetration. I prefer Pylyshyn's definition, that:

> if a system is cognitively penetrable then the function it computes is sensitive, in a semantically coherent way, to the organism's goals and beliefs, i.e., it can be altered in a way that bears some logical relation to what the person knows" (Pylyshyn, 1999: 5).

See also Raftopoulos (2005) and Rowlands (2005) for variations on how precisely (and at what level) to define cognitive penetrability.

[22] The McGurk Effect is often cited as *the* paradigm example of penetration, as it is a very robust effect that strikes even under fairly degraded perceptual conditions (see Sams *et al.,* 2005). Fodor, interestingly, doesn't

think that it threatens modularity, for either visual or parsing systems. He notes, "it is of central importance to realize that the McGurk Effect—though cross-modal—is itself domain-specific—viz., specific to language" (1983: 132, *n*13). We will return to this admission by Fodor in the discussion of the McGurk Effect later on.

[23] A study in a similar vein, by Gick & Derrick (2009: 502), found that "syllables heard simultaneously with cutaneous air puffs were more likely to be heard as aspirated (for example, causing participants to mishear 'b' as 'p')."

[24] Dyadic in the case of the purported instances of cognitive penetration listed above. As I will argue in chapters to come, integration systems can be more multivariate in terms of integration than merely dyadic fusing operators.

[25] It's worth noting in passing that Burnston & Cohen's position is that the "modularity" just needs to be definitionally refined in a fashion "recognizing that the modularity/non-modularity distinction comes apart from the cognitive/non-cognitive distinction, and that they together partition mental processes into four kinds rather than two. It allows for processes that are non-modular and cognitively penetrated (e.g., rational belief fixation); processes that are non-modular and not cognitively penetrated (e.g., prospective memory); processes that are modular and cognitively penetrated (e.g., mental arithmetic); and processes that are modular and not cognitively penetrated (the representation of chasing)" (Burstein & Cohen, 2014: 18). I disagree, for reasons that will be apparent below, though I will not return to their account specifically to rebut it in detail.

[26] Note that Carruthers actually refers to an "interaction *effect"* (emphasis mine)—outputs of various systems (perhaps as few as two, for example) could be "picked up" as the proprietary inputs to a further system, which merely integrates, or fuses them, or blocks one from proceeding to an output stage. That's an interaction *effect* – but not really the sort of *bidirectional* interactivity one imagines in cases of cognitive penetration, which is much more open-ended in implication.

[27] For an alternate take on this sort of cross-modal domain-specificity, see Aspeitia, A.A.B., Eraña, Á., & Stainton, R. (2010), where the argument is that the only genuinely useful construal of domain-specificity is one in which modules are viewed as operating over idiosyncratic representational domains (i.e., not over a common representational language of thought). The authors note that "this notion seems quite appropriate for input-output modules, but ... falls afoul of our third constraint: it does not apply to central mental systems" (26). I will argue extensively below that it *can* apply to central systems—much more on this in chapters 4-6.

[28] More on this idea of "compiled" modules in chapter 4. Marr's account was discussed in §3.1 above.

[29] Jackendoff's (2002) account of the McGurk Effect similarly posits an integration system between vision and hearing that is an example of virtually encapsulated "structure-constrained modularity":

> 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)

[30] This sort of account could support a reevaluation of so-called *motor theory of speech perception* (see Liberman & Mattingly, 1985, for a review), in which the assumption is that shared neural structures involved in both speech perception and production could serve mutual benefit—i.e., the ability to produce certain phonemes can help clarify perception, as the same structures are activated and past associations can fill in missing perceptual information, or resolve ambiguities. One possibility is that in the case of the McGurk Effect, the visual input of the sound production (i.e., the other's facial movements) could trigger, via mirror neuronal activations, a similar motor response rehearsal in the perceiver, thus activating the structures that help "figure out" what's being said. "Motor theory" is not limited to speech perception, either. Kiverstein (2010; *cf.* Gongopadhyay *et al.,* 2010) argues that all "contents of experience depend upon a perceiver's sensorimotor knowledge" (257) insofar as sensorimotor knowledge gives rise to expectations which let us experience whole 3D objects, exceeding what is available at any given time, also lends us perception of presence of intrinsic properties that we cannot perceive. More on this sort of idea below, when we turn to Barsalou (2009) and his simulation theory.

[31] It's not totally clear to me what Fodor would make of this sort of account. On the one hand, as noted above in note 24, he suggests the McGurk Effect is an artifact of modularity, as it is part of the language system, and domain-specific despite being cross-modal. But then, this suggests that a *cross-modal functional* module is not a problem, on Fodor's view. On the other hand, he clearly objects to this sort of account insofar as dismisses moves to functionally individuate modules: "the 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). In chapter 4 we will discuss Fodor's objections to this sort of account at greater length. For now, I will note my only my confusion as to how he thinks these two conclusions cohere.

[32] Note one seeming hole in this idea: how can we explain what happens when we shut our eyes (or just lip-read, with sound blocked)? How can we be consciously aware in *those* contexts, if my diagram is correct? Surely we *can* have direct conscious access to the outputs of the lip-reading stage, I would argue that it's *not* the case that the lip-reading module reports directly to awareness in that context: rather, the integration system still proceeds as normal—it's just that in the cross-check stage, there is nothing on one side to cross-check, so the data coming from the other sensory module gets a "free pass" through the system to the "integration" stage, and on to awareness. If you don't look at the screen in the McGurk video, the heard /ba/ gets integrated with *no visual information,* so it is processed *as is.* The important point to remember is that if *both* percepts are present, they *must* be integrated, like it or not.

[33] The list: *deal, eel, feel, meal, ... wheel, zeal.* Only one of these is going to work, based on the semantic understanding of the sentence (barring some new metaphorical usage). I suppose *seal* could work in some very small subsets of contexts. I would hazard a guess that *primed with* such a context, *seal* might show up in a phoneme restoration test. We will discuss *semantic priming effects* in much greater detail in chapters 5-7.

[34] There is an associated phenomena of *auditory priming*, in which what sounds like incomprehensible gibberish on one listen, can suddenly become sensible if you are *told* what it is saying (or simultaneously read what it says). Amazingly, once you hear it as sensible, i.e., not gibberish, you can no longer hear it as anything *but* sensible. We will hear what we are supposed to hear. This is fairly easy to manipulate as well, given the susceptibility people have to hear "hidden" messages in backwards speech, for example. If you are primed to hear it, and the percept is *close enough*, you will hear it, and you will never be able to *un*-hear it. The sensibility is non-overridable. This is a key element that we will return to in the discussion of non-revisable belief and delusion in Part III of this dissertation. For a good example, look for Michael Shermer's "Stairway to Satan" example from his TED talk "Why People Believe Weird Things", available online at <https://www.youtube.com/watch?v=8T_jwq9ph8k>.

[35] Another way of describing this is through an account of 'simulation' like that of Barsalou (1992; 2009). We will look at this idea in depth in §3.4, below.

[36] Metamers, are pairs, or groups, of surfaces that have distinct surface spectral reflectance distributions (different SSRs), but are nevertheless perceived as *the same color* by a given observer under the same viewing conditions. For more on metameric effects, and the puzzle they constitute for metaphysical theories of color, see Hilbert (1987), and Cohen (2009).

[37] The Necker cube (Necker, 1823) is ambiguous as to which face is to the front.

[38] Recall the Gick & Derrick (2009) study, noted in the previous section—in which subject who were administered a small puff of air to the neck were liable to "hear" non-aspirated sounds as aspirated ***b***s or ***p***s—a "simulation" theory could help explain how this relevant association could impose the aspiration on the signal. As Gick & Derrick conclude: "These results demonstrate that perceivers integrate event-relevant tactile information in auditory perception in much the same way as they do visual information" (2009: 502).

4 Framing beyond the periphery

In the previous chapter, I argued that the modularity thesis is the best way to explain how perceptual systems can tractably represent the outside world, despite the ubiquitous *frame problem,* by virtue of the informational encapsulation, domain-specificity and cognitive impenetrability of modules. I also argued that the evidence generally used against the modularity thesis—namely, evidence that purports to show cognitive penetration of ostensible modules, and violations of encapsulation—is actually better understood as evidence of modular, encapsulated, mandatory *integration* systems that refine and resolve perceptual representations both within systems and cross-modally across sensory systems. So far, the modular systems I have defended do not go much beyond the sensory periphery (at most, they draw modular boundaries around various compilations of lower-level sensory systems). In this chapter and the next, I wish to argue for modular cognitive structures that operate well beyond the periphery, at higher cognitive levels, and include processes that can underwrite concept acquisition, composition, action planning, deliberation, belief formation and revision. I argue that this thesis has *prima facie* plausibility based on (apparent) human successes in all of the above-listed cognitive activities—all of which run head first into the *frame problem*, and as we have seen, modularity of function, including informational encapsulation and domain-specificity are arguably the best way around that problem.

As we saw, modules, as defined by Fodor, are clearly immune to the frame problem insofar 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 courting the frame problem for the reasons stated above. The domain-specific encapsulated module has an innately specified frame, which allows for reflexively fast, always tractable computation. There is no question of modules getting bogged down by queries to and from other systems, since the module is cognitively impenetrable. Of course,

if modules 'solve' the frame problem at the most basic level; it's plausible to suggest might do so at *every* level. Since Fodor (1983), there have been numerous theorists who have argued this line, positing various higher-level modular systems—that the mind is more *massively* modular than Fodor argues for.[1] 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 independently from one another in terms of processing (for the most part). If all cognitive processes could be shown to be (in some sense) modular, this could allow for an account of how it's possible to maintain computational tractability in all aspects of cognition. However, Fodor himself believes that, ultimately, the move to employ modularity more massively is doomed as an answer, to the frame problem, or anything else:

> 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 (2000: 64-65).

Recall Fodor's self-professed "Quinean" attitude toward central systems—that belief revision *must* be "isotropic" insofar as "in principle, *any* of one's cognitive commitments (including, of course, the available experiential data) is relevant to the (dis)confirmation of any new belief" (2008: 115). Belief fixation and revision, in Quinean terms, needs to be holistic in order to ensure coherence and hence demands isotropy. Fodor argues that modularity, by definition, can't handle processes that are isotropic in this way. Encapsulation and isotropy don't mix.

> It seems clear that isotropic, Quineian systems are *ipso facto* unencapsulated; and if unencapsulated, then presumably non-modular . Or rather, since this is all a matter of degree, we had best say that to the extent that a system is Quineian and isotropic, it is also nonmodular. (1983: 111)

Similarly, systems that are domain-specific can't, by dint of architecture, have the access necessary to engage in belief formation and revision.

> We have repeatedly distinguished between what the input systems compute and what the organism (consciously or subdoxastically) believes. Part of the point of this distinction is that input systems, being informationally encapsulated, typically compute representations of the distal layout on the basis of less information about the distal layout than the organism has available. Such representations want correction in light of background knowledge (e.g., information in memory) and of the simultaneous results of input analysis in other domains ... Call the process of arriving at such corrected representations " the fixation of perceptual belief." To a first

> approximation, we can assume that the mechanisms that effect this process work like this: they look simultaneously at the representations delivered by the various input systems and at the information currently in memory, and they arrive at a best (i .e., best available) hypothesis about how the world must be, given these various sorts of data. But if there are mechanisms that fix perceptual belief, and if they work in anything like this way, then these mechanisms are not domain specific. Indeed, the point of having them is precisely to ensure that, wherever possible, what the organism believes is determined by all the information it has access to, regardless of which cognitive domains this information is drawn from. (1983: 102)

In this chapter, I will make the case that Fodor's pessimism about modularity beyond the sensory periphery is misplaced. I will present two examples of higher-level, post-perceptual 'assembled modules'—both socially oriented—in order to establish that modularity is almost certainly *more* massive than Fodor believes, and that a 'module', properly so-called, can indeed violate his initial criterion of 'non-assembly'. Additionally, I will briefly examine some proposals from Jackendoff (2002; 2007) and Carruthers (2006a), which attempt to trace out how modular processes might be able to explain even the most global cognitive operations, such as belief revision and abductive inference. The goal of this chapter is to establish the plausibility of the thesis that at least *some* assembled modules exist, and that belief fixation and revision can similarly be computationally constrained via modular systems of concept formation and associative memory retrieval processes. In the chapter that follows, I will build on these ideas, and further deflect Fodor's concerns about more massive construals of modularity, marshaling Fodor's own (1975; 1998; 2008) theory of conceptual atomism to support my thesis. For now, in this chapter, the goal is more modest: to simply argue for the existence of assembled modules, and to argue that Fodor's objections to modularity beyond the periphery are misplaced, and that, in fact, his own arguments lead inexorably to the conclusion that modularity *must* extend beyond the periphery, to at least some extent.

4.1 Assembled modularity

In the previous chapter, I already made a first pass at a definition of what an 'assembled' module could be in the sense of an *integration* system constituted by a compilation of sub-component perceptual modules. In this section, I am going to expand beyond mere perceptual integration systems to argue for the existence of assembled modules—modules built out of modules that can process higher level conceptual representations and *belief* in a

tractable fashion, taking advantage of the computational tractability benefits of modular architecture. Of course, to a strict Fodorian, "assembled modularity" is an oxymoron, as point #3 of his original (1983) definition of modularity explicitly decrees as a criterion of *non*-modularity that "the computational system [is] 'assembled' (in the sense of having been put together from some stock of more elementary subprocesses)" (Fodor, 1983: 36). A modular system must, on Fodor's definition, "map relatively directly onto its neural implementation" (*ibid*).

However, I actually think that this principle doesn't fit with Fodor's own conception of modularity. Charitably, I would assume that what he wants to rule out here are violations of encapsulation—and his criterion of non-assembly would protect encapsulation by ruling out inter-system accessibility. There are, of course, ways in which assemblies of modules might be described as interacting in a way that violates encapsulation—"cross-talking" and mutually penetrating one another. By ruling out modular assemblies, Fodor certainly rules out this possibility. However, I believe that his prohibition against modular assembly is much stricter than necessary to simply protect the virtue of informational encapsulation. It is not *necessary* that a modular assembly involve interactions that violate encapsulation and inaccessibility: a modular assembly could be arranged in such a way that the component modules are *not* mutually penetrable, but, rather, feed inputs to one another unidirectionally. Fodor wants to define modularity in such a way that penetration is impossible, and because *some* sorts of assembled modules could be construed as penetrable, he rules out assembly *tout court.* I am arguing that this is too hasty, and neglects the many ways modular assembly could be construed that do *not* invite penetrability. I think encapsulation is sufficiently upheld by the insistence that modular systems *not* "share horizontal resources (of memory, attention, or whatever) with other cognitive systems" (37). That does not rule out modular assembly of any kind, just modular assembly of a certain kind. What I do not see, in Fodor's argument, is a specifically compelling reason to deny unidirectional interaction between modular systems, as in the case that the outputs of one system are the proprietary inputs of the next system. We have already seen such systems proposed in the last chapter: cross-modal sensory integration systems. But we could go beyond mere sensory integration, and posit assembled systems that take on higher level concept composition or object detection roles—roles that seems to imply "thinking" and "knowing" all sorts of things, in order to

make *judgments,* but which, I will argue, could be simply the results of highly specialized conceptual integration mechanisms, with all the hallmark features of modular processors.

4.1.1 If it *looks* like a duck, *walks* like a duck, *quacks* like a duck...

As an example of what I mean, consider the following system hierarchy:

- MODULE A in activated by input stimulus ***a*** and outputs representation **α**
- MODULE B is activated by input stimulus ***b*** and outputs representation **β**
- MODULE C is activated by input stimulus ***c*** and outputs representation **γ**

Let's assume that stimuli *a, b,* and *c* occur regularly in the environment, and our hypothetical organism has evolved sensory apparati to represent them, associated with cognitive processing MODULES A, B, and C. Each MODULE—A, B, C—is informationally encapsulated, domain-specific, computationally autonomous and cognitively impenetrable, etc. So far, so good.

Now imagine a fourth module—MODULE D—which is tasked with a sort of *interface* function that operates over the *outputs* of MODULES A, B and C, and takes as input precisely (and only) representations of the form **α**, **β**, and **γ**. MODULE D is only activated when all 3 representations are input simultaneously.

- MODULE D, when activated by **α**, **β**, and **γ**, outputs **δ**.

Perhaps, for ease of explanation, we posit MODULE D as a "duck detector". MODULE A picks up on any stimulus that looks like a duck*;* MODULE B does the same for things that *walk* like a duck; and MODULE C detects *quacking* sounds. MODULE D takes the outputs of these three as inputs, and only outputs *THERE'S A DUCK!* in the case that sub-modules have simultaneously converged on *looks like a duck, walks like a duck* and *quacks like a duck.* Arguably, such a MODULE could emerge if there were selective pressure to be extremely accurate at responding to ducks, but *not* to sort-of-duck-like entities, impostor ducks, swans, platypuses, etc.

This MODULE D story is perfectly plausible, at least from a computational or architectural standpoint—indeed, we will look at proposals for a number of similar sorts of perceptual integration "modules" in this chapter. The key takeaway is that such a MODULE

D could still be said to meet all the criteria of modularity, specifically and most importantly, the criteria of informational encapsulation, cognitive impenetrability and domain-specificity of function. The only "violation" of traditional Fodorian modularity is that it is *assembled*.

Furthermore, if my MODULE D story is unconvincing, note that if Fodor is right that systems "put together from some stock of more elementary subprocesses" can't count as modules, then *vision* shouldn't count as a module either, as should be obvious from the discussion of vision in the previous chapter. The visual 'system' is an *assembly* of transducers—or a "compilation" as Pylyshyn (1999) would describe it, each tasked with extracting a different value from the distal stimuli—whose outputs are coordinated and processed in stages before output as representations. The transducers exhibit encapsulation of function (they have proprietary inputs, and shallow delimited output stages) that essentially constitute *one level of framing*. The vision *module* then takes the outputs of transducers, imposes regularization algorithms—assumptions—on the transduced data, and outputs the visual representation. This is essentially a *second level* of framing.[2]

We could run the same sort of argument with regard to the language "module" that Fodor certainly believes exists: whatever a language "module" would be, it would surely be an *assembly* of subcomponent modular systems (phonological parsing, syntactical parsing, binding, etc.). Indeed, one quite influential theory of sentence parsing is Lyn Frazier and Janet Fodor's (1978) two-stage "sausage machine" account:

> That the syntactic analysis of sentences by hearers or readers is performed in two steps. The first step is to assign lexical and phrasal nodes to groups of words within the lexical string that is received; this is the work of what we will call the Preliminary Phrase Packager, affectionately known as the Sausage Machine. The second step is to combine these structured phrases into a complete phrase marker for the sentence by adding higher nonterminal nodes; the device which performs this we call the Sentence Structure Supervisor. These two parts of the sentence parsing mechanism have very different characteristics, and this provides an explanation for the relative processing complexity of certain types of English sentence. The Preliminary Phrase Packager (PPP) is a 'shortsighted' device, which peers at the incoming sentence through a narrow window which subtends only a few words at a time. It is also insensitive in some respects to the well-formedness rules of the language. The Sentence Structure Supervisor (SSS) can survey the whole phrase marker for the sentence as it is computed, and it can keep track of dependencies between items that are widely separated in the sentence and of long-term structural commitments which are acquired as the analysis proceeds. (Frazier & Fodor, 1978: 291-92)

If this is right, then again, just as with vision, the "module" in question *is* an assembly of subcomponents, computationally autonomous with respect to one another, whose outputs are

subsequently integrated and processed. In short, I think even at the level of relatively uncontroversial, Fodorian, modularity, we find systems that could be plausibly described as assemblies—modules built out of sub-modules. I see no compelling reason to deny that a "modular" system—properly so-called—could not in principle be built out of component modules. And there could be many, in nested hierarchies (INTERFACE MODULE D takes outputs from MODULES A, B, C; and INTERFACE MODULE Z takes outputs from MODULES X & Y; and SUPER-MODULE **ψ** takes as its inputs the outputs from Z & D.)[3] Or there could be multiple interface devices that operate over different, but overlapping, subsets of outputs from lower-level modules.

Indeed, this view seems eminently plausible: that nested hierarchies of modular systems render *all* cognitive processing tractable by subdividing task domains, and that even global, central systems may be decomposable into constituent dedicated processors. Two main arguments militating against this are (1) evidence of purported cognitive penetration, and (2) Fodor's defining of modularity as specifically excluding modular assembly. I have attempted to rebut (1), insofar as the evidence offered for cognitive penetration *supports* the massively modular view rather than creating problems for it. And I have suggested that (2) is an overly strict response to the penetrability question—Fodor's concern is misplaced and overlooks ways in which modular assembly could be construed to maintain encapsulation, inaccessibility and domain-specificity. Inputs are inputs—if the "module" in question meets all the other criteria of modularity, then why is it a deal breaker if the input to that module *just happens to be the output of another module?* That doesn't seem to me to be a violation of encapsulation—the "higher" level module isn't really *accessing* the lower module—rather, the "higher" level module may just be adapted to picking up the (shallow) outputs of certain other modules. And it certainly doesn't seem to violate domain-specificity: indeed, the domain of the "higher" level interface module may be far *narrower* in many instances than the lower level sensory modules that feed it. In the next 2 sections, I want to look at two proposed modular assemblies: a *theory of mind module (ToMM)* and a *cheater detection module (CDM)*. I will not be making a committal argument as to the existence of such modules; I merely present them as models of how assembled modular structures could plausibly function.

4.1.2 A theory of mind module

The folk-psychological ability to attribute intentional and epistemic states to others is one of the hallmarks of human cognition—and one which arguably appears to meet at least *many* of the Fodorian criteria for modularity: it seems reflexive, mandatory, quick, developed along a characteristic ontogenetic course, and subject to systematic breakdown.[4] Simon Baron-Cohen (1996) presents an elegant modular account of a theory of mind faculty comprised of two perceptual-level input modules (an *Intentionality Detector* and an *Eye Direction Detector*), which interface through a *Shared Attention Mechanism*, and are assembled (or compiled) in an overarching *Theory of Mind Module,* which develops according to specific ontogenetic schedule, and may be prone to systematic patterns of breakdown.

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.[5] After the *ID* comes online, a normally-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 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: representations of eyes or eye-like entities.[6]

According to Baron-Cohen, a bi-domain-specific interface 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 so Mommy cannot see the object" (Baron-Cohen, 1996: 67).[7] 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 contention 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 (Baron-Cohen, Leslie, Frith, 1985).

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 attribute 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).[8] *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). 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 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.[9] Rather, I highlight his account here to underscore the point 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, and its individual processing is encapsulated at the local level. *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* and the capacity for pretense. 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.

4.1.3 A cheater detection module

A 'cheater detection' (or 'free-rider' detection) module may seem like a bit of a stretch, though it has been argued for at length by evolutionary psychologists Leda Cosmides & John Tooby. The motivating principle is clear: being capable of quick, correct representation of the motives (especially *ill*-motives) of one's conspecifics is highly important to evolutionary success, so it is *prima facie* plausible that adaptive pressures would select for cognitive functions that could facilitate that task. If it's plausible that we have evolved an innate system to detect *literal snakes*, it seems equally plausible that we have one for *metaphorical snakes:* deceptive people who are out to take advantage of us. Many empirical findings on human reasoning suggest that there are unconscious biases toward faster and more effective processing of logical operations, for example, when the content is social in nature. A famous example is the Wason selection task performance effect, in which people reason *better* and more logically in social contexts (Wason 1968; Cosmides, 1989).[10]

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:

- **Cards with a vowel on one side, must have 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:

A	K	4	7

Most subjects recognize the need to turn over the card with the vowel—they affirm the antecedent—but many fail to logically determine the need to turn over the odd-numbered card—they neglect to deny the consequent! 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 simple logic puzzle, and the result is no surprise: people generally are quicker to employ *modus ponens* than *modus tollens*. But the interesting data comes when the abstract logical relationships between symbols and the given rule is substituted by more meaningful, concrete, *social* 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	Car	Train

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). Suddenly, more people remember to deny the consequent in their checking procedure. This 'content effect' was viewed as initially quite puzzling,

given 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 & 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 & 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. 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, operates subdoxastically, and its algorithm is apparently not generalizable for use in other logically equivalent situations. We are simply better as logical reasoning when it involves socially-oriented content in which deception or "cheating" is a live possibility—but that logical "skill" is non-transferable to other reasoning task, and it is completely opaque to us (i.e., we do not have any inkling that we are reasoning any better or worse as content shifts).[11]

4.2 Fodor's challenge

Fodor, as we have seen, is not on board with extending modularity beyond the sensorium, and explicitly rules out these sorts of "assembled" modules. At this point it is worth bringing in his main, *a priori* objection to the project of more *massively* construing modularity: what Fodor (2000) calls the *input problem*. The input problem is fairly simple: 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). Figure 8 illustrates the problem:

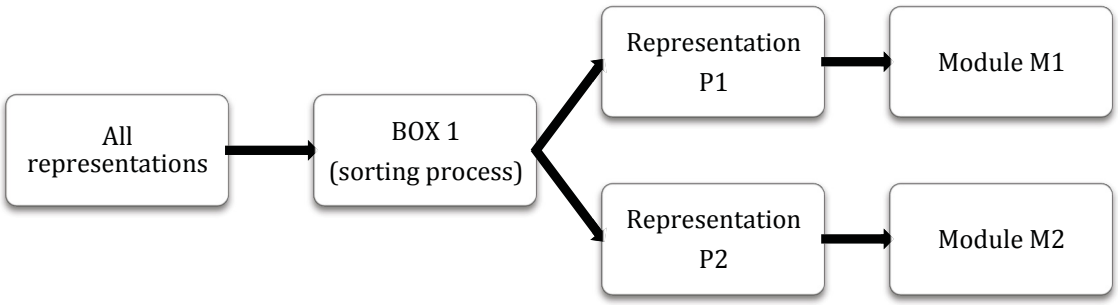


Fig. 8: Fodor's "input problem". Recreated by the author.

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 process (**BOX 1)** that handles the sorting and assigning. Fodor's point is that this **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 **BOX 1** which can take in *all representations* and begin the assignment 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). Frankish (2004:

58) calls this "Fodor's challenge"; Carruthers (2003) calls it "Fodor's problem". Whatever we call it, there does seem to be a lurking question as to how dedicated, post-peripheral, post-perceptual cognitive modules are to be supplied with the requisite inputs. Carruthers, espousing a massively modular account, argues that natural language faculties are the key:

> 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).

On this account, *everything* feeds the language module, which can then decompose and recompose conceptual inputs via the massive recursive resources available to language production. This on the surface seems plausible enough—the mere fact that we are able, generally, to "vocalize" (internally or externally) what we are experiencing, feeling, doing, etc. suggests that the language production faculty has some sort of access to the outputs of the systems that take care of processing all that experiencing, feeling, doing, etc. What Carruthers seems to argue from there is that this internal "speech" can then be (re)run through perceptual and/or motor systems to provoke *new* content, essentially. But Fodor's problem still seems to be a fair problem, on Carruthers' account, for two reasons: a) this description of the "powers" of the language production system doesn't explain what *directs* those powers—it seems that Carruthers has just thrown the whole mystery into the language box and is suggesting the magic happens there; and b) how does requerying the perceptual or motor system get relevant outputs any closer to an appropriate module, like say the *CDM?* Fodor says there needs to be a domain-general sorting between perception and the further module—how does re-querying the perceptual module, solve that? Wouldn't the re-iteration merely lead us back to BOX 1?[12]

With respect to the *CDM*, Fodor takes aim at it specifically, and at Cosmides & Tooby's arguments concerning its ostensible encapsulation. 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, 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).

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

> it 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 (2000: 78).

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

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

Note Fodor's reintroduction of the frame problem regarding abductive or holistic inference: his argument here is that even just in the consideration of *one* post-peripheral modular assembly (a cheater detection "module"), we have no choice but to posit some kind of executive "agent"—essentially a sorting device—that has access to *everything* in the cognitive background, in order to send inputs to the appropriate processing "modules". And if that is the case, the modules don't do us any favors—the framing problem they are invoked to solve just shows up at the input stage.

However, I think Fodor is mistaken on this point, both in the case of the proposed *CDM,* and on the broader challenge of his "input problem". Whatever input problem Fodor imagines strikes at the heart of post-peripheral modules should be equally (or more) problematic at the sensory periphery as well. I'm not sure how he thinks *vision* is any better off in this regard. Vision, surely, picks up on "very subtle cues" in order to represent depth, edges, color, motion, etc.—indeed, *so* subtle in most cases that "cerebration" would not even be capable of working them out. *And that's the point:* modular processing evolves because it (just so) happens to pick up on something usable and do something useful with it. Fodor compares the detection of social exchange to the language system, suggesting that solving the input problem for language processing is at least *maybe possible* since there are "psychophysical" cues in the stimuli when it comes to language, whereas in detection of social exchanges there is not. But I see no reason to make this distinction, for two reasons: **1)** Why *aren't* there cues regarding "social exchange" contexts that are "psychophysically detectable"? Think of what needs to be "detected": I would argue that the cue to a "social exchange context" is merely the *presence of a conspecific,* which is a problem of visual categorization and therefore a problem of input. A "social cognition" system could be *activated* by the perception of a conspecific in the environment. What the system *does* with that information, upon activation, is debatable—I'm not really that concerned with making an argument one way or the other on that score—my point is simply that the "input" problem doesn't seem like much of a problem. Conspecific detection alone could be enough to engage a "social" module, or more plausibly, a whole suite of distinct "socially-oriented" processors that sift the data for different elements and process accordingly, one of which

might be a processors dedicated to precisely the "very subtle cues" Fodor mentions. If a handful of "very subtle cues" are *the only thing* a certain module "sees", then it's not surprising if that module sees them.

2) Even if one wants to stick with Fodor's contention that the "very subtle cues" of social contexts are too variable and contextual to be detectable by (dumb) modular processors, if *language processors* are capable of unpacking the "very subtle cues", and language processing is itself a prior, modular system, then your input problem to the "social" module (or modules) may evaporate, at least in the more narrow cases where the social exchange "cues" are in the form of verbal interaction with another (or hearing testimony of such an interaction, or *reading* about it, or merely *imagining* it, in the sense of an interior narrative). Once in the language processing system, one need only assume some automated processes of parsing and semantic interfacing via which data can be tagged as "social exchange" and then served up as input for the "social" module. Of course, a Fodorian would flag the "semantic interface" stage as being unframed. I have not made any argument (yet) as to how that process might (itself) be encapsulated, and not require access to the entire epistemic background. As a promissory note, that is to come later in this chapter and the next. For now, let me just suggest that the process of attaching meaning to language certainly has all the hallmarks of a modular process: surely we don't *think* about what words mean too much, in ordinary communication.

Again, whether or not the account above is correct is not crucial to the central argument I want to make in this dissertation. All of the proposals above are merely plausible seeming possibilities that suggest Fodor's challenge, as we have called it, doesn't seem quite as devastating as he suggests it is. I believe the challenge can be met, whether or not any of the assembled modules proposed and discussed above turn out to exist or not. The "input problem" is not the sweeping *a priori* objection Fodor makes it out to be, although it may still be a very serious problem for a modular account of global, holistic belief revision and inferential reasoning capacities. As for Fodor's prohibition against assembled modules *tout court*, there is not a clear principle that rules them impossible, and I think that I have provided evidence that at least some are plausible and likely. But that's not enough to get us to the goal of a *complete* account of tractable belief revision. Fodor is right to continue to insist that in order to do that, we need an answer to the frame problem that can account for

abductive inference—as belief revision will be "up to its ghostly ears in abduction" (Fodor, 2000: 75). As mentioned earlier, Fodor is a self-professed "Quinean" in the sense that belief fixation, confirmation, revision all must be construed as *isotropic*, insofar as "what the organism believes is determined by all the information it has access to, regardless of which cognitive domains this information is drawn from" (1983: 105). To respond to this, it's not going to be as simple as proclaiming a *belief revision module (BRM)* will take care of it—since such a module, in order to access all relevant belief in every instance, would need to give up encapsulation and domain-specificity, and would hence lose all claim to modularity.[13] On this point, Fodor is correct. The point where I think he is *incorrect* is in his strict Quineanism about belief. I don't think the "central systems" that manage and mediate belief revision need be isotropic or Quinean in the way Fodor describes. Indeed, as we discussed in chapter 1, the Quinean requirements on belief revision—coherence, conservatism—are, at best, regulative ideals. *No* finite computational system could plausibly meet those norms.

The upshot is that one can't seemingly be *both* a Quinean and a computationalist. Fodor's answer to this is to simply deny computationalism when it comes to central systems. My answer will be the opposite: we will have to accept that whatever the "central systems" that mediate belief are, they aren't Quinean. What we have, instead, are a suite of modular subsystems that employ heuristic search and processing strategies to *approximate* the sort of isotropic, Quinean system Fodor favors. Much more on this in chapter 6. Though before we get to that point of my account, I will develop a few lines of argument against Fodor in chapter 5 which will hopefully help make the case in chapter 6 more intuitively plausible.

4.3 Review and look ahead

So far, we have looked at a few proposals for assembled modules that operate in a fashion that *appears* global in nature, with access to a seemingly broad domain of background belief and/or semantic knowledge, and yet can arguably be modeled in a way that is highly domain-specific and encapsulated as to processing. We have also looked at Fodor's "input problem" with regard to such systems, and I have argued that his concern is overblown, and that inputs in the form of representational output by lower-level systems could be proprietary to certain post-perceptual processing modules. In the next chapter, I want to try and extend the idea of

assembled modular structures to an account of how belief is tractably managed—in a way that is *not* Quinean, as Fodor believes it should be, but rather fully constrained by frugality constraints imposed at various levels. To get there, I want to engage Fodor's objections a little more—and invoke his own theory of conceptual atomism, and his story of concept acquisition, as expressed in his two books on the *language of thought* (1975; 2008) to show that his own account dovetails pretty nicely with a more massively modular account than he is ready to admit. Chapter 5 will looks at a number of aspects of concept acquisition, storage, retrieval, and the processes that takes us from percept to concept to belief. I will show that even if we make the same initial assumptions that Fodor makes, the deliberative processes that are engaged in order to form and fix and revise belief do *not* have access to the entire epistemic background, and there are ways to model how they can still effectively arrive at and subsequently revise belief without that unrestricted global access. Two ideas will be essential to tell this story: 1) how heuristic approximations can substitute for more Quinean, holistic belief revision practices; and 2) a 'global workspace' theory of working memory that explains how only severely restricted (task-specific) domains of information can be retrieved and processed at a given time. These are the central ideas will be explored in chapter 6.

Notes for chapter 4

[1] Some examples of massively modular accounts are Barkow, Cosmides, & Tooby (1982), Carruthers (2006a), Sperber (2005), Barrett & Kurzban (2006). All of these will be referenced later on in the chapter.

[2] Indeed, we saw in the previous chapter that Pylyshyn (1999: 5) notes that there are commonly "within-vision effect[s] — i.e., visual interpretations computed by early vision affect[ing] other visual interpretations, separated either by space or time."

[3] "Super"-module only in the sense that it is farther down the processing stream—it need not be "bigger" in any sense. Indeed, as we saw in the discussion of Jackendoff (2002) in the previous chapter, it's quite possible that interface (super)modules could actually be *more* domain-specific and thereby more limited than lower-level feeding modules, as they operate over sub-sets of the data processed by at lower levels.

[4] It would be beyond the scope of this thesis to really examine those claims in detail, so in this section I will only sketch Baron-Cohen's (1996) account and gesture towards the claims to modularity. See Baron-Cohen (1996), Baron-Cohen, Frith & Leslie (1985) Segal (1996), Andrews (2005), Carruthers & Smith (1996) and Siegal & Surian (2006) for more discussion of the possible modularity of theory of mind.

[5] 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.

[6] 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. Further support for the modular, encapsulated function of "gaze detection" can be found in Friesen & Kingstone (1998); Hood *et al*. (1998); Deaner, Shepherd & Platt (2007).

[7] I.e., they understand the concept of occlusion, despite never visually experiencing occlusion. This could also serve as further support for the sort of interpolation module discussing in the previous chapter. We apparently don't need to "learn" about occlusion via experience. We might also connect this observation to a classic philosophical puzzle about perception: Molyneux's problem, as presented to Locke, regarding whether a blind man who knew globes and cubes by touch would be able to recognize them by sight alone if his sight were suddenly restored (Molyneux, 1693). Locke, as an empiricist should, responds in the negative (Locke 1693). I would be inclined to think the prospects are more positive, myself.

[8] 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 discussion to follow regarding "cheater detection."

[9] Not everyone agrees with Baron-Cohen's account of the *ToMM*. Prinz (2006) disputes the supporting data:

> Consider another example: massive modularists claim that we have an innate capacity for "mindreading," i.e., attributing mental states. The innateness claim is supported by two facts: mindreading emerges on a fixed schedule, and it is impaired in autism, which is a genetic disorder. Consider these in turn. The evidence for a fixed schedule comes from studies of healthy Western children. Western children generally master mindreading skills between the third and fourth birthdays in normally developing children. However, this pattern fails to hold up cross-culturally (Lillard, 1998; Vinden, 1999). For example, Quechua speakers of Peru don't master belief attribution until they are eight (Vinden, 1996). Moreover, individual differences in belief attribution are highly correlated with language skills and exposure to social interaction (Garfield et al., 2001). This suggests that mindreading skills are acquired through social experience and language training... What about autism? I don't think the mindreading deficit in autism is evidence for innateness. An alternative hypothesis is that mindreading depends on a more general capacity which is compromised in autism. One suggestion is that autists' difficulty with mindreading is a consequence of genetic abnormality is oxytocin transmission, which prevents them from forming social attachments, and thereby undermines learned social skills (Insel et al., 1999).

[10] 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; Tooby, Cosmides, Barrett, 2005; Cheng & Holyoak, 1989; Clarke, 2004. The original publication of the experimental results can be found in Wason, 1968.

[11] This non-transferability of the additional skill is highly relevant, both to the argued modular status of this case, and also later on in this thesis, when we look at pathological belief states which are often *not* logical in any sense, and yet logical abilities seem intact with regard to other beliefs the subject has. This shows that some "judgment" and belief-generating processes can be more or less dissociated form logical reasoning, and that logical reasoning is not an "all-purpose" tool that can be accessed equally in cognition. Much more on this in PART III.

[12] There is additionally a problem lurking in Carruthers' account (which I will return to again, below, as Jackendoff (2007) critques this aspect of Carruthers as well)—namely, the question of how sentences produced by the language faculty (inner speech) can be "displayed in motor imagination". It's not at all clear how the motor system "imagines" anything, nor is it clear how natural language sentences could serve as inputs to motor control, without some other mediating processes.

[13] Mercier & Sperber (2009; 2011) argue that there might be such a belief revision module, in the sense that there is could be a modular "argument analysis" system:

> What the argumentation module does then is to take as input a claim and, possibly, information relevant to its evaluation, and to produce as output reasons to accept or reject that claim. The workings of this module are just as opaque as those of any other module, and its immediate outputs are just as intuitively compelling. We accept as self-evident that a given pair of accepted assumptions of the form

> P-or-Q and not-Q justifies accepting the conclusion P, but this compelling intuition would be hard to justify. (Mercier & Sperber, 2009: 9)

I will not take up the task of critiquing this idea here, as I am not quite clear on how they imagine this module to work, in practice. If it is simply a module that can identify "arguments" (at least those in standard forms, i.e., "if...then...") and evaluate them via *modus ponens* and perhaps *modus tollens*, then this may well be part of the bigger picture of a modular belief revision system. I don't think it's likely to be as simple as Mercier & Sperber describe, however. The "argumentation module" is likely more like an interaction effect of multiple levels of parallel modular processing, rather than in individual module.

5 Concepts, belief, and Fodor *vs*. Fodor

In this chapter, I want to work out some arguments rebutting Fodor's contention that Quinean considerations militate against construing belief fixation and revision as modular processes. Ironically, I intend to employ a number of Fodor's own ideas about concept formation, storage and retrieval in order to make that case. In short, I will argue that Fodor's understanding of concept acquisition requires the involvement of numerous post-perceptual modular assemblies, despite his objections to such an extension of his account. The refrain of Fodor's anti-modularity argument regarding belief centers on the global/local distinction—specifically, the problem posed by the *locality* of computational processes for any explanation of more *global* human cognitive capacity.[1]

> Computation is, by stipulation, a process that's sensitive to the syntax of representations and nothing else. But there are parameters of beliefs (and hence, representational theory of mind being assumed, of the mental representations that express those beliefs) that determine their role in non-demonstrative inference but are, on the face of them, not syntactic: relevance, conservatism, simplicity are extremely plausible examples. So either learning, belief fixation, perception, and the like aren't processes of non-demonstrative inference (but what on earth else *could* they be?) or they aren't computations. The upshot is that the more a mental process is plausibly not local, the less we understand it (Fodor 2008: 124).

Fodor's line is firm: belief fixation and belief revision *cannot* be modular, computational processes precisely because they go beyond syntax, and they require global access to the epistemic background. This task can't be met by local, domain-specific, encapsulated mechanisms. On the surface, the objection seems perfectly intuitive: the entire point of invoking modularity at lower levels of processing is to *impose limits*. But belief formation and revision are ostensibly *un*limited—especially once we are in the realm of scientific theorizing and abductive inference: how could you possibly do something like *find a 'best' explanation* to a problem using a system predicated on limitations of processing? Belief has to be Quinean and isotropic, as every belief is potentially relevant to every other belief, and hence must be accessible to it. But Fodor says: "modularity is fundamentally a matter of informational encapsulation and, of course, informationally encapsulated is precisely what Quineian, isotropic systems are not" (1983: 110). Of course, the entire discussion in this

dissertation was predicated on the assumption that we really simply *can't be* Quineans: that even in those most "global" realms of cognition, we *do* require limits—the upshot of what Cherniak calls the *finitary predicament*. Furthermore, the specter of the frame problem demands that our account of cognition *must* include some mechanisms to limit even the seemingly unlimited processes of belief revision and abductive inference. We can't be Quineans, computationally. Fodor's response to the problem is that we *must* be Quineans about belief, so this should simply lead us to drop computationalism, at least beyond the peripheral systems that are plausibly modular. I have been arguing for the opposite conclusion: keep computationalism, and accept that belief isn't Quinean, and isn't isotropic, properly speaking, after all—it is something more constrained than this, though which at times can *mimic* or approximate the Quinean ideal.

Cain (2002), though he is largely sympathetic to Fodor's account, finds Fodor's rigidity on this point to be confusing. He wonders why Fodor is so reluctant to admit the possibility of "the central system's having a modular structure", since it seems to fit perfectly well with Fodor's account that "the central system decomposes into several distinct sub-systems each of which has a distinctive function which it executes by running its own specialist program. One such system [might be] the theoretical reasoning system" (Cain 2002: 199). I agree with this point: Fodor's reluctance to invoke modular processes in (putatively) "central" processing seems misplaced, given so many of his other commitments. If concepts are acquired largely automatically and subdoxastically, and stored in highly organized way—as Fodor's theory of concepts contends—then it seems highly plausible to suggest that relevant associations between concepts (and hence beliefs) can be tracked and evaluated by way of massively parallel associative processing, instantiated by modular substructures, rather than global, Quinean central systems. In short, I am suggesting that Fodor's concern about employing modularity at the level of so-called central systems (which manage belief) is not only mistaken—it doesn't even cohere with Fodor's own account of concepts. In this sense, the goal of this chapter is to fight Fodor with Fodor.

5.1 Modular concept acquisition

In this section, I want to make the case that concept acquisition is directly mediated by modular processes. I have already argued that sensory processing is fairly uncontroversially

modular, and I have tried to defend an account specifying multiple levels of post-perceptual cross-modal integration modules that refine perceptual scenes as well as highly specified evolved detector systems to pick up on salient cues in the environment. In addition, I have argued that it is extremely plausible that multiple modules could "work together" as assemblies in the sense that some modules are dedicated to pick up outputs of others for further processing, or to re-activate lower levels in recursive iterations. In this way, we can get what amounts to a circumscribed, overarching "module" in the style defended by Carruthers (2006a), and other massive modularity theorists—where the modular assembly is what Carruthers refers to as an "interaction effect" of modules working in parallel such that a sort of "virtual" overarching module emerges. The resulting system can still claim to be "modular" insofar as the various modules in aggregation exhibit a *de facto* informational encapsulation and domain specificity, as the processing loops may be quite narrowly circumscribed.[2]

The position I wish to defend in this chapter is one that describes the process of concept acquisition as entirely mediated by modular assemblies, and hence we may say that concept acquisition is "modular" in this sense. This is not to say that there is a "concept acquisition module"—there is not, in my view: such a module really would beg the *frame problem*, as it would be domain-general in the extreme, tasked with constantly surveying the entire epistemic background. However, there is plausibly a concept acquisition *process* which is subserved by modules and exhibits all the telltale characteristics of modular functioning (no matter what level one inspects it at, from the level of individual concept, to the birds-eye view of the "system" as a whole). Fodor of course objects to any such elaboration and expansion of the modularity thesis, for reasons already discussed. However, in this section, I will try to show how Fodor's theory of concept acquisition *demands* modular processing to make it work. Fodor (2008) resorts to a somewhat poetic metaphor to "explain" concept acquisition, strangely reluctant to employ the better, more concrete explanation that fits exactly the same bill: modularity.

5.1.1 Fodor's LOT and conceptual atomism

To begin, we need to look at a bit on background of Fodor's theory of concepts, including how it operates through the language of thought (LOT). Fodor's (1975) *Language of*

Thought, though highly influential in cognitive science, and credited with helping spark the 'computationalist' program in general, is not the first to posit the existence of some sort of 'language of thought' or *Mentalese*.[3] Any plausible argument in favour of a representational theory of mind (RTM) implies a need for some sort of syntactic system in which representations can interact causally with one another according to systematic chains of (typically inferential) operations, forming, in essence, a type of mental language.[4] Additionally, as far as Fodor is concerned, LOT brings with it the bonus of getting around many troublesome 'Frege cases' of coreferentials in opaque contexts—the standard move of dissociating of *sense* from *reference* turns out to be superfluous on Fodor's account. Take the paradigmatic case of Cicero and Tully: we can clearly imagine situations in which a person who does not know that Cicero=Tully can have beliefs about one, but not the other. However, if, as RTM insists, mental representations have causal powers, then how can one form different beliefs about coreferential concepts with identical contents? The Fregean story has it that we must chalk this sort of concept individuation up to *senses*, or what Fodor refers to as modes of presentation (MOPs). For Frege (Fodor argues), MOPs must be extra-mental, in order to ensure that they can be public. But, as Fodor argues in *Concepts*, this is incoherent insofar as "if MOPs are to individuate mental states they will have to be the sorts of things that the causal role of a mental state can turn on. But it's a mystery how a MOP *could* be that sort of thing if MOPs aren't in the head" (Fodor 1998: 21). As to exactly what sort of thing the MOPs might be, Fodor's explanation is a little clearer in *LOT 2,* where he suggests that Mentalese offers a way out, insofar as it is computational.

> *CTM slices mental states thinner* than mere [propositional attitude] psychology does [...] CTM distinguishes the causal powers of mental states *whenever* they are tokenings of type-distinct mental representations, *even if the semantic contents of the representations tokened are the same"* (Fodor 2008: 70).

The upshot of this is that a person can very clearly have two different syntactic formulas tokened for separate (but coreferential) concepts—for example, one Mentalese formula for CICERO and another for TULLY. With this in mind, it becomes simpler to explain how these concepts could play different roles in *belief* despite having the same referential content, since the representations, being different formulas in Mentalese, have distinct causal powers, and may enter into different relations with other representations. Note that this also helps explain the pervasiveness of at least one form of the *inconsistency awareness* problem,

regarding how a person might have numerous contradictory beliefs and yet be unaware of them: Lois Lane's beliefs that Superman can fly and Clark Kent cannot are not directly inconsistent *to her*, as these beliefs are formed via the concepts SUPERMAN and CLARK KENT, which despite being coreferential, Lois "knows" under different MOPs. She has separate mentalese formulas for each, and they interact with belief separately as a result. According to Fodor, everything that senses are deemed necessary for in order to explain concept individuation can be achieved much more easily by simply assuming a computationalist framework and letting the MOPs individuate concepts at the syntactic level—not as something extra-mental, as Frege would have it, but internally, as different formulas in Mentalese. Fodor notes:

> In the long run, computational psychology is a sort of trick that Turing invented to make it seems that there are senses and that they cause things (even though, strictly speaking, there aren't and (therefore) they don't). The rule of thumb: if there is something that it seems that you need sense to do, either do it with syntax or don't do it at all (Fodor 2008: 87n).

One objection that this kind of answer might provoke (already hinted at above as Frege's possible motivation for invoking extra-mental senses) is that this sort of entirely internalized concept individuation would violate one of the main criteria of what a concept *has* to be (number 5, according to *Concepts),*[5] which is that concepts must be *public,* at least if successful communication regarding those concepts is going to take place. The objection to Fodor's LOT here would be that by relegating MOPs to different Mentalese formulas, his account violates the publicity constraint and makes it an utter mystery how two people could *share* a concept. Fodor notes this objection in *LOT 2*, and answers it by invoking a "dual-role" analogy of Mentalese tokens as "files" and also "file *names"* in order "to explain how formulas of Mentalese can play both these roles; how its formulas can both apply to things in the world and causally interact with one another in the course of mental processes" (Fodor 2008: 93).

> In a nutshell: Tokens of M(John) can function both to refer to John in our thinking and to interact causally with tokens of other mental representations in the course of mental processes. That's because mental representations can serve both as names for things in the world *and as names for files in the memory*. I want to pursue this file metaphor. (Fodor 2008: 94)

A concept "file" can contain various "memos" linking it to other (associated) files. For example, if John is your brother, then the file M(John) will contain a "memo" linking it to

the file M(brother). And if you think your brother John is a jerk, then presumably the file for John also contains a "memo" connecting it to the file M(jerk); and of course, the file M(brother) will contain M(John) and the file M(jerk) will contain both M(John) and M(brother), etc. Note that in this way, Fodor is describing our *beliefs* about various concepts as memos in the conceptual file. Your belief that John is your brother (and your belief that he is a jerk) are to be found in the file named "John". Thus, merely accessing the file facilitates access to all associated concepts (and belief).

> In effect, according to this story, *we think in file names*; tokens of file names serve both as the constituents of our thoughts and as the Mentalese expressions that we use to refer to the things we think about. If you are given John's name in Mentalese, you are *thereby* given the Mentalese name of a file where you keep (some of; see below) what you believe about John. That one thinks in file names is the best short summary I'm able to formulate of the version of RTM that I'm currently inclined to endorse. (2008: 95)

With the file metaphor that Fodor has in mind, we get a story that explains how concept *association* can, in principle, take place in a tractable way. Fodor uses the example of thinking HOUSE—with the file name M(house)—which may associate with the file named M(window), since HAS WINDOWS is part of the descriptive content carried in the M(house) file, insofar as stereotypical houses have windows. This allows for the kind of "semantic priming" effects that let us think associatively (when doing so is relevant).

> So you can get from M(house) to M(window) faster than you can get from M(house) to M(fish). That is plausibly A Good Thing since, quite likely, you will want to move from M(house) to M(window) faster than you get from M(house) to M(fish). (Fodor 2008: 98).

But the filename idea *also* explains why all associated content *isn't* automatically brought before one's mind, as standard associationist accounts might seem to imply. Fodor argues that "[a]ssociationists hold that associative bonds cause regular co-tokenings of mental representations. But then, since everybody knows that typical houses have doors and windows, why doesn't everybody think *door* or *window* when he thinks *house*?" (2008: 106). Fodor's story is that if we construe concepts as files, then it is not the case that every associated concept will be "brought to mind" along with any given concept. In order to bring an associated file to mind you (literally) need to search for it by name. The point is that associated files will involve shorter and quicker searches, as the "memo" taking one to the associated file is accessible within the file one currently has in mind. This notion of being able to move between some "files" more quickly than others is a crucial factor in the

account—one which I will exploit later on in this chapter. In short, I argue that this exactly the sort of "compartmentalized" storage system Cherniak suggests is demanded by finite systems attempting to approximate global search procedures: it offers an elegant solution to the question of how searches of the epistemic background can be rendered tractable. Fodor will not agree with this analysis, of course. I just want to note here that this is the avenue I will pursue in later sections of this chapter. But before we get to what I will argue is a strikingly *modular* concept organization structure, I need to first explain in some more detail how concepts are acquired and organized into file structures, according to Fodor's (2008) account—an account that centers on a peculiarly poetic image of "whirlpools" in a cognitive "attractor landscape".

5.1.2 Attractor landscapes—whirlpools of the mind?

In *Concepts* (1998), Fodor lays out the basics of his account of informational semantics, arguing that our primitive (atomic) concepts are informationally constituted by nomic mind-world relations. He employs 'appearance properties' to explain the metaphysics of how concepts refer to the items in the world to which they are locked. Using the concept DOORKNOB as his example, he explains that the concept DOORKNOB is constituted by the locking relation between that representation and the property of *doorknobhood* in the world—precisely in the way that GREEN is constituted by being locked to the mere 'appearance' of the property of *greenness*, which is not inferred, or projected as a hypothesis on green things. Fodor notes that this analogy to 'appearance properties' is the only way around the objection to inferential role semantics that it should be *mysterious* how we lock to the appropriate concept only when faced with the corresponding items in the world that exhibit that property (the so-called "doorknob/DOORKNOB" (d/D) problem that our concept of DOORKNOB should turn out to be tokened only via experience with doorknobs, rather than something else, like rabbits).[6]

Previous attempts to make sense of this have relied on the suggestion that perhaps to have a concept is to have a *stereotype*, which would help explain how the concept then corresponds directly to the appropriate objects in the world that prompt the formation of the stereotype in the first place. It could be that via some process of inductive or statistical inference, we intuit a hierarchy of dominance and sisterhood relations holding between items

of experience, and we subsequently apply our concepts to items in the world based on where in the hierarchy individual items fit. Fodor recognizes that there is a clearly intuitive connection between stereotypes and concepts, insofar as "concepts really *ought* to be stereotypes":

> Not only because there's so much evidence that having a concept and having its stereotype are reliably closely correlated (and what better explanation of reliable close correlation could there be than identity?) but also because it is, as previously noted, generally *stereotypic* examples of *X-ness* that one learns *X* from (1998: 138).

However, Fodor rejects the stereotype theory on the grounds that stereotypes violate compositionality, whereas concepts do (and must be able to) compose.[7] He uses the example of the complex concept PET FISH: anyone who possesses the (presumably atomic) concepts PET and FISH can construct and understand the concept of what a PET FISH might be. But if we want to argue that concepts are actually stereotypes, then we won't be able to make sense of this: according to Fodor, to know the stereotype of PET and the stereotype of FISH does not assure that one will be able to understand the concept of a PET FISH (Fodor 1998: 102). His point here is that for most of us, something like a *goldfish* is the stereotypical "pet fish"—but it seems highly unlikely that a goldfish would be our stereotypical example of a pet fish if we had never seen (specifically) a pet fish. If a child knows what a PET stereotypically is and knows what a FISH stereotypically is, it is not likely that these two concepts can immediately be composed into something that converges on a gold-fish-like concept. Fodor & Lepore (1996) explain:

> Prima facie, however, the distance of an arbitrary object from the prototypic pet fish is not a function of its distance from the prototypic pet and its distance from the prototypic fish. In consequence, knowing that PET and FISH have the prototypes that they do does not permit one to predict that the prototypical pet fish is more like a goldfish than like a trout or a herring, on the one hand, or a dog or a cat, on the other. (Fodor & Lepore 1996: 263)

Indeed, one might be inclined to agree with Fodor on this point and think that if one were to work from the stereotypes of PET and FISH to compose a stereotypical PET FISH, one would be more likely to land at CATFISH or DOGFISH than GOLDFISH, based on lexical association alone. Furthermore, concepts cannot be stereotypes, according to Fodor, because if they were, it would make a mystery of how we gain concepts at all—for if we learn a concept X via the experience of STEREOTYPE OF X, then how is it we end up with the concept X, rather than the concept STEREOTYPE OF X? "What you'd expect people to

reliably learn from stereotypic examples of X *isn't the concept X but the X stereotype* (Fodor 1998: 138-139).

So, it can't be stereotypes that our concepts lock to, according to Fodor—instead, we are left with the somewhat obscure explanation that we simply have "the kind of minds" that just so happen to "lock" or "resonate" to certain "appearance-like properties" in the world:[8]

> My story says that what doorknobs have in common qua doorknobs *is being the kind of thing that our kinds of minds (do or would) lock to from experience with instances of the doorknob stereotype.* (Cf. to be red *just is* to have that property that minds like ours (do or would) lock to in virtue of experiences of redness). Why isn't that OK? (Fodor 1998: 137).

Why isn't that OK? Well, for many critics it just seems implausible that our minds should be so designed as to lock onto things like *doorknob-ness*. It's plausible to suggest that there are syntactically described representations innately predisposed to lock certain concepts to properties of objects in the world that seem central to adaptive success (like snakes, perhaps, echoing the arguments earlier in §3.1.3). *But doorknobs?* Putnam, for one, argues that the entire idea of an innately endowed syntactic system that gives rise to "innate semantic representations" for such things as doorknobs and carburetors is evolutionarily implausible on the face of it, since it suggests that "evolution would have had to be able to anticipate all the contingencies of future physical and cultural environments. Obviously it didn't and couldn't do this" (Putnam 1996: 15).

In *LOT 2*, Fodor has a new analogy for the locking process. Gone is the more elusive description that we simply "have the kind of minds" that lock or resonate to certain properties in the world in such a way as to constitute the content of primitive concepts. In the newer analogy, Fodor describes our innately endowed cognitive architecture as an "attractor landscape" full of "whirlpools".[9] I'll first let Fodor explain his idea, and then I'll try to paraphrase what he means by this in more concrete terms:

> The mind is like a sea [...] Imagine a sea that's dotted with boats, all sailing along, as happy as larks [...] There is, however, a catch. Randomly distributed over the sea on which the boats are sailing, there are whirlpools [...] into which things may fall according to the principle that the closer to a whirlpool a thing gets, the greater the force with which the whirlpool tries to suck it in [...] Think of concepts as attractors, each with its location in the sea. Think of stereotypes as boats in the sea located according to the principle that the better the stereotype, the closer it is to the corresponding attractor [...] And the closer a stereotype is to an attractor, the more likely it is that learning the stereotype is sufficient for acquiring the concept, that is, for locking to the property [...] Get close to an attractor and you lock to a property. That's

> a brute fact about the kind of animals we are; and it's the bedrock on which the phenomenon of concept acquisition rests (2008: 159-161).

So what are we getting with this picture? In short, this analogy gives us a condensed explanation of how stereotypes mediate concept acquisition, without constituting conceptual content. In *Concepts,* as we have seen, Fodor makes the case that our concepts cannot *be* stereotypes, since stereotypes violate compositionality, yet stereotypes nevertheless suggest themselves as having a somewhat ambiguous role in the story of concept acquisition because it is only through experience with stereotypical *f*s that one's mind can lock to the property of *f-hood.* In the 1998 account, this was left a bit mysterious and just chalked up to our 'having that kind of mind'. In *LOT 2*, this attractor landscape analogy is meant to highlight in more detail *how* the experience with stereotypical *f*s gets us to lock on to *f-hood*—stereotypes *can* be learned, or generated via statistical inference, and when a stereotype is learned 'properly', then it gets close to an innate whirlpool programmed to suck in the stereotype. At this point, a Mentalese formula is tokened and assigned to this new concept (we 'open a file'), the contents of which are the properties of items in the world that the stereotype was formed to be a stereotype of. This explains how it is only through experience with *f*s that we lock to *f-hood* and form the concept *F*, and that the *f*-stereotype we form is not merely a by-product, but is actually a necessary, though insufficient, stage on the way to concept acquisition.

Fodor schematizes the process of concept acquisition as in Figure 9:

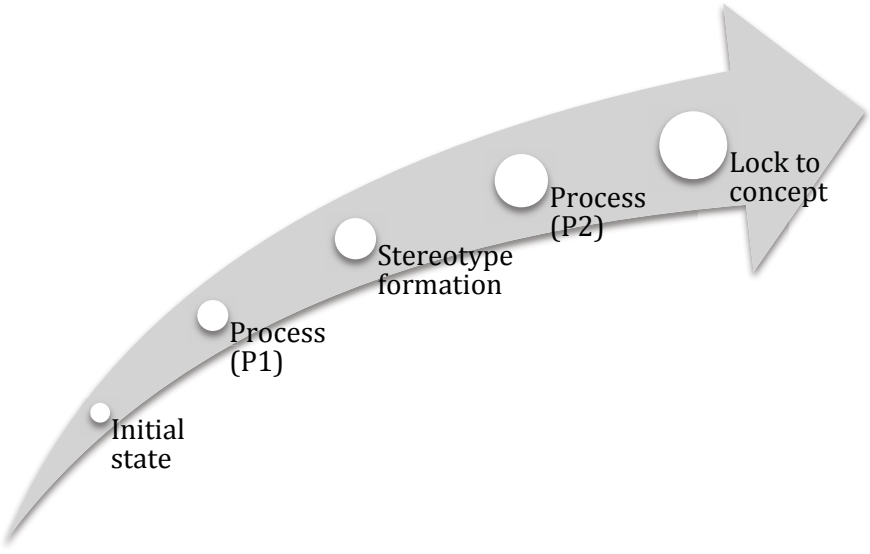


Fig. 9: Fodor's concept locking process. Adapted by the author, from Fodor (2008: 151).

In this progression, process (P1) is some process of statistical inference (perhaps innately instantiated and entirely subdoxastic, or (alternatively) trained up and employed consciously—it doesn't matter for Fodor's purposes). He notes that "even very young infants are able to recognize and respond to statistical regularities in their environment. A genetically endowed capacity for statistical induction would make sense if stereotype formation is something that minds are frequently employed to do" (2008: 153). (Note that this is an important admission, and one that will play a crucial role in the arguments I will make in the following sections.) After stereotype formation, process (P2) is some reliable but *non*-intentional, non-inferential neurological process—"a *sub*intentional and *sub*computational process; a kind of thing that our brain tissue just does" (Fodor 2008: 152). In simpler terms: we form a stereotype for some X via statistical inference based on repeated exposure to items in the world that are stereotypical examples of X. Once we have formed, or 'learned' the stereotype in the traditional psychological sense, our neurology takes over, and the attractor landscape drags appropriate (well-formed) stereotypes into locking relations, or resonances, with properties in the world, and a concept is attained. The concept is *not* the stereotype, but the stereotype mediates the locking in of the concept, and is a necessary stage on the way there. In some instances the stereotype will be so close to its corresponding whirlpool that for all intents and purposes it will seem identical to the concept, but this is still not to say that the stereotype *is* the concept.

Let's try a concrete example of concept formation to try and make sense of Fodor's attractor landscape account: what goes on as a child learns the concept DOG? On Fodor's picture, we are to imagine that the child first forms a stereotype of *doghood*, either by noticing a brute statistical pattern in nature, or by having the stereotype "trained up" by active ostensive definition, guided by parents and others. Children point, and inquire as the names of things, and try out those names, getting continually praised or corrected in their usage. At some point, a child may *appear* to have a concept of DOG, as she is capable of successfully and consistently recognizing instances of dogs *as dogs*. On Fodor's view, of course, the child may not have fully "locked" onto the property of doghood, but merely the stereotype. One test for this is the extent to which the child may *overgeneralize* the stereotype of doghood to things that aren't, in fact, dogs. Indeed, anyone who has ever watched children learn to employ lexical concepts will notice that children often

overgeneralize the concepts they have learned, and must fine-tune them—for example, having successfully "learned" the concept DOG, a child may, upon seeing a lamb or a calf for the first time, call it a "dog". Presumably this happens because the lamb or calf conforms closely enough to the stereotype of doghood that it seems a fair enough fit. Indeed, this is precisely why children's books (and parents) cross-culturally focus on the *sounds* various animals make as a way to differentiate them conceptually via metonymic associations. A lamb looks enough like a poodle to count as a dog, perhaps, *but it doesn't bark.* And dogs don't say "baaaaa". So the child's overgeneralization of the concept DOG to an instance of a lamb will be revealed as faulty, and will be revised. On Fodor's account, when the stereotype of doghood that the child has formed is sufficiently close to including all and only the properties of *dogs* (rather than, say, sheep or cows), the concept will "lock", and a file is opened in mentalese for DOG, in which all the properties of doghood known to the child are filed.[10]

5.1.3 Modular "whirlpools"?

It's clear that a lot of the foregoing account may seem like mere hand-waving about what *might be* the case, all hinging on what many would regard as a simply loopy idea of these cognitive *whirlpools* sucking in representations to lock them to conceptual content. Fodor himself seems to accept that his whirlpool analogy is largely a poetic way to simply try and explain what he ultimately views as a mysterious fact about a "kind of thing that our brain tissue just does" (Fodor 2008: 152). I think Fodor actually has a much better, less mysterious explanation at his disposal, though he doesn't accept it: *the whirlpools are modules.* Atomic concepts, represented in mental syntax, get picked up for processing via various modular processors—not sorted ahead of time, not 'knowingly' sought out by homuncular-esque modules—the 'whirlpools' are just dedicated processors that are matched with certain content, and process that content in a way that 'fixes' a concept and outputs it for filing in memory. That's it. Fodor describes the concept locking procedure as something that happens automatically, beneath consciousness, according to a specific ontogentic program, and (given that each whirlpool is dedicated to the acquisition of a singular concept) is strictly domain-specific and encapsulated. In short, concept acquisition, according to Fodor himself, is subserved by an attractor landscape that has all the markings of a suite of massively

parallel modular processors. Of course, we could reintroduce Putnam's objections here: that it is evolutionarily implausible that we would have a concept-framing module in place for DOORKNOB or CARBURATOR. I don't think it needs to be that overly specified though—perhaps simply modules to frame basic, highly adaptively useful concepts such as TOOL or COMPONENT. From there we just subdivide files in that folder based on *other* associations—at this point we might as well consider the concepts *definitions.* That doesn't seem evolutionarily implausible at all.

So, to be clear: I am suggesting that the "attractor landscape" idea sounds like a metaphor for a suite of modules, dedicated to forming particular concepts as well as sub-concepts formed by mining previously held stereotypes for conceptual parts. As discussed in chapters 3 and 4, above, humans have plausibly evolved highly specified modular "detectors" for things like *snakes* or *eyes* or *intentionality* or even *cheaters*—it doesn't seem much of a stretch to think that many more basic concepts, including things like DOG and ANIMAL, and natural kind concepts, for example. This basic level of concept acquisition would require one layer of individual concept acquisition 'devices'. From there, we will need mechanisms that can tractably form stereotypes, which can then mediate further concept learning. For example, the DOORKNOB concept won't be a directly acquired basic concept, since it's implausible that we have an evolved "doorknob detection" system. But what we *could* have is a set of detectors for the more basic concepts and stereotypes that mediate the acquisition of DOORKNOB—i.e., insofar as doorknob are just a very specific sort of TOOL. DOORKNOB will certainly be a later addition to the conceptual store of items that can be classed under TOOL, but the function of the doorknob is so specific that it could individuate naturally with repeated perceptual experiences, and lock-in as a free-standing atomic concept constituted by *doorknobhood*. This doesn't require a specific module on the lookout for *doorknobs*—it requires a module on the lookout for "things that fall under the stereotype and are not already included under other concepts": an organization module, one might call it. Rather than the robustly nativist position Fodor has, where we have whirlpools each individually set to lock to a *particular* property, it's much more likely that we have processes set to lock to *some (any)* property. If it turns out to be a property that once subsumed under a concept does useful work, it will stick around in the cognitive economy. If it is redundant it

can be discarded, or merely absorbed by another. And if it is useless, it will be ignored and forgotten—nothing will ever activate it.

Additionally, if stereotype formation is, as Fodor contends, largely a result of some sort of unconscious statistical inference generation, then, again, it sounds like he is talking about something that has all the hallmarks of being modular: a dedicated processing algorithm that sorts perceptual stimuli, runs pattern recognition, and feeds a statistical inference engine. The outputs of that system are stereotypes, which then can be picked up by relevant concept acquisition modules (or sucked into the whirlpools, as Fodor has it). From this, we can get to a full stock of atomic concepts, each locked-in via an "attractor" module and "filed" in memory. Again, to stress the point about what I am *not* arguing, I am not suggesting that there is a module for every concept. It can't actually be the case that individuated modules exist singly for each concept, both for the Putmanian reason discussed above, and further, if it *were* the case, then we should expect to find all sorts of people who fail to ever grasp one or more single individual concepts precisely because of some very small bit of brain damage (damage to the "whirlpool" in question). And yet, to the best of my knowledge, there is no empirical evidence of people out there who are simply *incapable* of acquiring a single, specific, concept, while capable of acquiring other concepts normally—if there is a problem with concept acquisition, it is a more general problem, not a problem with *specific* concepts.[11]

Furthermore, if modular processing underwrites concept acquisition as I am arguing, we should expect the concept formation and composition to exhibit some telltale signs of modularity: specifically that it is automatic, domain-specific, encapsulated, follows a predictable ontogenetic path and is prone to systematic forms of breakdown. And in fact, this is exactly what we seem to find in the initial stages of lexical concept formation in children. One empirical data set that we can turn to is what, in the connectionist literature, is commonly referred to as the "U-shaped" acquisition curve that children exhibit in language acquisition (Bowerman, 1982; Karmiloff-Smith, 1979,1986; Pinker & Prince, 1988, Plunkett & Marchman, 1991). This is the phenomenon (well known to anyone with small children) where, for example, irregular past tense verb forms are learned and successfully used by a toddler, and then when the rule for *regular* past tense construction is learned by the child, she *over*uses it, and begins to add *–ed* to all past tense verbs (*goed, eated,* etc.), despite having

previously used the correct (irregular) form (*went, ate).* As Plunkett & Marchman (1991) explain,

> overgeneralizations typically occur after children have been using correct forms of irregular verbs appropriately. With development, the organization of the linguistic system supports the correct production of both regular and irregular past tense forms. This apparent regression and subsequent improvement suggests that acquisition involves a stage-like reorganization of rules and representations and is an oft-cited example of U-shaped development. (Plunkett & Marchman, 1991: 44)

The idea here is that at first children have learned the verbs individually—the past tense of each verb is essentially stored separately in memory. Of course, if this were to continue, it would be a massive waste of cognitive resources, and make search times impossibly long as vocabulary expanded. So the system learns the "rule", based on the most regular pattern: add *–ed* for past tense. At which point *all* verbs get subsumed under that rule, and presumably the previously individually stored lexical concepts are re-mapped (or re-filed, depending on your chosen metaphor) under the rule-concept. There will then be a period of time in which the child systematically makes a mistake with the irregular verbs, until correction and training can introduce the secondary rule: the rule for irregular exceptions.[12]

Granted, Fodor is not a connectionist, so he may or may not be happy with a comparison between his "whirlpool" concept-acquisition schema and the discussion of U-shaped acquisition, though to my eyes, they seem to be describing very similar processes—processes via which concepts are provisionally fixed via stereotypes, and later reorganized under various hierarchical rule-governance structures (concepts about concepts). Using the file-structure analogy, we can see how concepts, as they are acquired, are installed and organized in multiple levels of overlapping indices. The upshot of this is clear: concept acquisition is a largely self-organizing process that, crucially, sets up concepts in a structure that is conducive to later searches that maintain tractability via associative processing. Finally, I want to highlight the many aspects of these processes that seem to display exactly the characteristics one associates with modular functioning: the acquisition follows a clear ontogenetic path, it is not accessible or introspectible, it appears to be mandatory, and it is prone to systematic breakdown—the "misfires" (such as overgeneralizations) follow clear and predictable patterns.

Of course, the topic ostensibly under discussion in this dissertation is not *concept* formation, but, rather, *belief* formation (and even more so, revision). So why am I spending

all this time trying to show that Fodor's theory of concept acquisition is arguably modular? My main motivation is that belief formation and revision are conceptual processes: as Fodor himself argues, concepts are the constituents of beliefs (2008: 25). Of course, the formation of belief involves *doing something* with those concepts, so there is an additional process going on than mere concept formation. My contention is that this *secondary* process is already made tractable by the modularity of the acquisition and organization of the concepts which constitute beliefs: the idea is that if concepts are formed via modular processing, there will be an attendant organizational structure which allows for tractable search and associative processes *between* concepts thereby facilitating subsequent composition and revision of concepts and the beliefs they constitute. In the next section I will elaborate on this idea, and in §5.3, I will complete the connection between concept formation and belief formation as modular processes.

5.2 Modular concept organization

Most of this chapter is dedicating to "fighting Fodor with Fodor" as I described it in the introduction: trying to defuse standard Fodorian objections to the employment of massive modularity as a way to explain tractable belief management by showing how other elements of Fodor's theory of concepts actually support the massively modular view. However, in this section, I will add a bit more non-Fodorian support to this claim, specifically with regard to how concepts seem to be organized in memory and in associative relationships that allow for tractable search and revision processes. I will look at three distinct, but complementary arguments in this section. First, I will explain why I think Fodor's account of concepts lends itself to the thesis that concepts are *self-organizing*, in an entirely unconscious, subdoxastic fashion, purely as a result of their being acquired via parallel modular processing. Second, I will compare this view to Barsalou's theory of conceptual storage in recursive "conceptual frames" to highlight how this latter account helps to ensure the tractability we want to preserve for a system of concepts (and from there, beliefs). Finally, I want to look briefly at some of the work of Endel Tulving on memory encoding and retrieval, which similarly highlights the associative organization of concepts in memory, and crucially includes empirical evidence that leads to the next step of the argument: that belief formation, insofar

as it involves retrieval of concepts from memory, is *not isotropic* and is *not Quinean*, in practice, despite Fodor's insistence to the contrary.

5.2.1 Self-organizing concepts

Recall that on Fodor's account, concepts are organized in file structures, and are tokened individually via their filename (which is a formula in mentalese). That formula serves a dual role however: it is both a filename (that can be actively searched, and can enter in causal (syntactic) relations with other mentalese tokens), and it is also a file that contains organized information (including relevant associations to other files, and beliefs about those files). Fodor's account of how these files are accessed and searched demands a global executive—indeed even the subsuming of files within files seems to demand one (i.e., who or what is writing those "memos" in the files? Fodor is committed to some central executive function that can *access all the files* and understand the semantic connections between them). Yet, I want to deny precisely such an executive. I want to argue that the files should automatically self-organize as concepts "lock" in place, on a strict interpretation of Fodor's account. If we take the attractor-whirlpool idea seriously, then we are accepting that every individual concept was, at some point, "sucked in" to a whirlpool once a close-enough stereotype was formed with associated content. But we also are meant, on Fodor's account, to presume that stereotypes can serve multiple duty in multiple instances of concept formation, and that more primitive concepts can be composed into more complex concepts etc. These too, then, should require a whirlpool to lock them in. But then, this suggests that as ever more complex concepts are formed, they will lock-in some associations *automatically and subdoxastically*—specifically, those associations that can be shared by individuated concepts that were mediated by the same stereotype (in full or in part). So the organization of conceptual files is no less "a kind of thing that our brain tissue just does" (Fodor 2008: 152) than the formation of initial, atomic concepts. There is no central, global executive or central system that needs to assign associations and write memos into files: the files should simply self-organize with multiple, overlapping associations intact, inherited from the stereotype that mediated acquisition of the concept. And by extension, I would argue that subsequent searching need not rely on any global central executive—automatic associative processing can get from concept A to concept Z relatively quickly if Z is filed under A. Recall Fodor's

point about the filename structure: that such a filing system makes it faster and easier to move from the concept HOUSE to the concept WINDOW than it is to move from HOUSE to FISH – which he calls "a good thing" given the more relevant association between houses and windows.

Of course, Fodor suggests that associative processing can't be the general story, as it would clog up cognition and "one's thinking would be forever getting derailed by one's associations" (2008: 96). However, I don't see why this would follow. Let's take HOUSE and WINDOW as an example: suppose I am in a context where I need to visit a friend, whose home I have never been to before, and all I know is that his is "the third house on the left after the bend in the road ahead". Fodor, arguing against associationism, asks an interesting question: "since everybody knows that typical houses have doors and windows, why doesn't everybody think *door* or *window* when he thinks *house*?" (2008: 96). Clearly, in practice, surely I can think about "the third house on the left" without thinking of windows at all. However, all this suggests is that the association doesn't rise to the level of consciousness in that context—which is not to say that it isn't *activated*. Indeed, I would guess if the third building on the left after the bend were a windowless concrete bunker, I would immediately be puzzled and wonder "could this be the house?" *precisely because* the lack of windows undermines the connection of this building to the concept HOUSE I was operating from. Why? Because, regardless of whether I was conscious of the association on the drive up, the association between WINDOW and HOUSE was so automatic and obligatory that WINDOW immediately became salient as a missing but (unconsciously) expected item. I was looking for a HOUSE, but apparently I was also looking for a WINDOW, without explicitly setting out to do so—the associated concept *was* present to my mind without consciously being so. This speaks to the strength (and unconscious automation) of the associative processing that is a function of conceptual storage—and it is precisely the sort of indicator of modular functioning that I am arguing in favour of.

In fact, so far everything I am describing about the *self*-organization of concepts suggests modular processing: it is automatic, fast, follows a developmental path, it is neurologically instantiated ("the kind of thing our brain tissue just does"), domain-specific, informationally encapsulated (to the extent that whirlpools do their organizational work impenetrably, regardless of context), and prone to systematic breakdown (i.e., U-shaped

acquisition curves, obligatory associations, etc.). Everything Fodor is committed to at this ostensibly "central" level of organization seems amenable to an account that assumes the central system decomposes into constituent modular parts, yet Fodor resists the idea. In the next section, I want to leave Fodor aside for a moment to look at Barsalou's idea of "conceptual frames" which I will argue is largely similar to Fodor's file-based conceptual organization, and offers many insights that further support my modular account of concept organization.

5.2.2 Barsalou's 'conceptual frames'

We have already looked briefly at Barsalou's (1992; 2003; 2009) account of "situated conceptualization" in chapter 3—an account that describes how perception is the result of a multi-modal integration via "simulation". Barsalou contends that perception is streamlined and sped up in its processing by the activation of memories of previous experiences with relevant feature associations/similarities—the memory will automatically activate a "simulation" from which expectations can be generated, and subsequently compared to the incoming percept in order to smooth out irregularities, attenuate noise in the signal, rule out error, and interpolate where there are gaps, etc. We looked at that account with reference to modularly-mediated sensory perception, but it's worth coming back to Barsalou here, in chapter 5, to see the next step of the process from perceptual smoothing to the firming up of learned concepts that result from perception, and the organization of those concepts in what Barsalou calls "frames". Barsalou's conceptual frames feature a sort of "recursive embedding" (1992: 162) of atomic conceptual parts (features, attributes, whatever one might refer to them as).

> Consider the frame for *car*. Each of its attributes is actually a more specific frame: *Engine* is a frame with attributes for *ignition system, fuel system, lubrication system, cooling system*; in turn, *ignition system,* is a frame with attributes for *battery, starter, distributor;* in turn, each of these attributes is a frame and so forth. (1992: 162)

Frames can assist with concept learning, as frames can overlap, and attributes that comprise one concept can partially comprise another as well. It can also help explain what we refer to as *dispositional* or *tacit* knowledge or understanding of concepts: to use an example from Fodor (2008) we have tacit knowledge that Shakespeare did not have a telephone. Why? Because we can quickly infer that from the information we already have in our cognitive

store: SHAKESPEARE includes GUY WHO LIVED IN RENAISSANCE TIME PERIOD; PHONE includes INVENTED POST-RENAISSANCE. Logic takes care of the rest. Barsalou can explain it easily with conceptual frames: the frames for PHONE and SHAKESPEARE simply don't overlap in any way, as they are embedded in a historical/temporal situation frame that has them separated in a way that cannot be activated together. However, this does still leave us with a question regarding what organized these recursive embedding of frames. How are they stored and organized, exactly?

How to *store* concepts poses a tricky question, the answer to which depends on where limitations lie elsewhere in the cognitive system. On the one hand, the most economical storage design in terms of *space* would be an *inheritance* model, where "each property is only represented once, at the highest level for which it is generally true, yet it is *inherited* by all concepts along any descending chain of *type* relations" (1992: 178). For example, all animals eat—so for any given animal, a dog, for instance, the fact that it eats does not need to be explicitly stored within the concept of dog. The concept DOG would simply inherit the concept EATS. (And so would CAT, and COW and all other animals—EATS need not be stored again and again in each case). Now, we certainly *could* do it this way, especially if concept mediation proceeds mediated by stereotype or prototype formation as argued above: we could learn DOG and BARKS and FUR and MAMMAL and ANIMAL and EATS and all the conceptual relations between those, and store them in a treelike *type* relation structure. "Thus *eats* is true of every subordinate concept that descends through *type* relations from *animal.* Consequently, the properties true of *dog* include *barks* (directly associated), *fur* (inherited from *mammal)*, and *eats* (inherited from *animal)"* (*ibid*). However, Barsalou suggests that although this is an economical and elegant storage solution, it would make searching and processing far too demanding:

> The lack of cognitive economy in human knowledge demonstrates an important trade-off between storage and processing. Cognitive economy in representation optimizes storage, because categories and properties are not stored redundantly. Optimizing storage in this manner, however, incurs high processing demands: to find all the concepts and properties true of a concept, it would be necessary to search up its type chain and accumulate inherited information" (180).

So, Barsalou concludes, a redundant storage strategy is preferable—despite taking up more space in one sense, the processing will reach its halting point faster. So *storage* economy is made up for in *search* economy. This highlights one of the major conditions Cherniak

suggested with regard to the *finitary predicament:* compartmentalization of memory in order to expedite searches. Better to store concepts compartmentalized (packaged) with the relevant features, rather than demand one process up the type chains to find all the inheritance relations that apply.

This sort of proposal seems directly analogous with Fodor's "filename" approach to concept organization, and again has many features suggestive of modular processing. One in particular that Barsalou notes is that repeated activations of certain conceptual associations can lock in the association to the degree that an associated concept may be *obligatorily* activated, regardless of contextual relevance. Most of the time the context *will* be relevant, which is why the association is generally strong—but there will be times when the association is not contextually relevant, and yet it is activated automatically regardless. We will see in the chapters to follow that this is an inevitable side effect of this sort of associative conceptual or cognitive priming: often features of a concept that are irrelevant to the processing task at hand are nevertheless called up and processed—corrupting the process in many instances. We will examine the many sorts of "cognitive illusions" that can arise as a result in chapter 6. Barsalou highlights the Stroop phenomenon (Stroop, 1935) as a good example of this:

> Does the information included in a conceptualization depend completely on context? Several investigators have found that some properties are included across all contexts, regardless of whether they are relevant... Recall the Stroop phenomenon... as we saw, the meaning for *orange* is activated obligatorily, thereby interfering with naming purple as the ink color... To the extent that a particular property is associated with a word consistently across many contexts, its activation becomes obligatory, regardless of whether it is relevant in the current context. (1992: 180)

Obligatory activations due to repeated use (or usefulness) are an excellent time-saver (as well as an indicator of underlying modular processing). The cost is non-optimal concept management, distortions and corrupted processing due to the presence of competing yet irrelevant associations, but presumably the result is satisfactory enough in most cases.

5.2.3 Memory encoding and retrieval as modular processes

One of the primary theories of memory in psychology—that of Endel Tulving (1972; 1983)—also describes a storage and retrieval system that is optimized for tractability via associative priming and a similar sort of recursive embedding and storage organization as

described by Barsalou. Tulving's account is directly relevant to the question of tractable concept and belief revision for reasons that I will examine in this section: in short, the way memories are recalled (including beliefs stored in memory) turns out not to be amenable to an isotropic or Quinean reading, such as Fodor gives.

Tulving explains memory storage as being quite unlike what most people intuitively assume it to be: remembering is not simply a fetching of recorded information. For Tulving, memory is not *recalled,* so much as essentially *reconstructed*, making use of associated items that are activated in semantic memory that help interpolate and smooth out the many disparately encoded bits of information that are tied to the "event" being remembered. We generally speak of memories as something "stored" discretely, and subsequently activated through retrieval. On Tulving's view that's not quite the case: he suggests most theorists who talk about memory are prone to a "storage bias" (1991:7). But what is stored, according to Tulving, is a distribution of encoded data in two distinct, but interacting systems: an *episodic* memory system, in which the perceptual data from temporally tagged events are encoded, and a *semantic* memory system in which semantic knowledge (rules, definitions, algorithms) are stored, which interact with data in the episodic system during acts of retrieval.[13] These are *engrams*, as Tulving calls them, borrowing the term from Semon (1904; see Schacter, Eich & Tulving, 1978, for a review). On its own, an engram does not constitute a memory, however.[14] The available engram has to be joined with a "retrieval cue" in a process Tulving calls *ecphory*: "the process that combines the information in the engram and the retrieval cue into ecphoric information [which] determines recollective experience, the end product of an act of cognitive memory" (1991: 6). The remembered information has to be essentially reconstructed according to the retrieval cue in *synergistic ecphory:*

> Ecphory is one of the elements of episodic memory, a component of the process of retrieval; 'synergistic' refers to the joint influence that the stored information (the engram) and the retrieval information (the cue) exert on the construction of the product of ecphory… for a complete understanding of the 'what' and the 'how' of retrieval we must also take into account the substantive contribution made by non-episodic information present at retrieval. (Tulving, 1983: 12)

The "substantive contribution made by non-episodic information present at retrieval" refers to the fact that the recall context can actively shape the substance of what is recalled. On one hand, this can be very useful insofar as it gives a head start in maintaining search tractability

(given that the search context can shape—and thereby limit—the result of the search, automatically). But it also can give rise to a number of memory distortion and (false) belief perseverance effects that I have discussed throughout this dissertation, and will discuss further below, in chapter 7. This is the price of a constructivist picture of memory, as the reconstruction may not go exactly as the original encoded event may have gone—the context of retrieval may be different than the context of encoding. Roediger (2000) explains that "retrieving is like perceiving for a sentient observer" (72) insofar as the act of (re)constructing a memory in *synergistic ecphory* is itself an event for the rememberer, and hence can affect a change to the episodic system in the process of accessing engrams within that system. Allik (2000: 16) similarly points out that there are multiple ways to extract information from internal representations—depending on the retrieval context, one's memory may be quite different from one instance of retrieval to another. Memories are not stored as discrete units, but rather are stored as a distributed set of parts, which can be composed into a discrete memory, but can be variously mixed to some degree, and their composition may be highly sensitive to the recall context (though we won't be consciously aware of this). Furthermore, each instance of retrieval, as it is a sort of "reliving" of the experience, becomes *itself* another episode to encode (i.e., you can remember that time you remembered that time...). The upshot of all this is that memories can essentially be reshaped via (unconscious and automatic) associative processing—but given that beliefs are also stored in memory, we are already getting a hint as to how one might argue that associative processing can similarly reshape and revise belief.

But is Tulving's view plausible? What evidence is there to support the view that memory is constructed occurently, and doesn't technically exist in "storage" as common usage and folk theory suggest?[15] Numerous studies appear to support the idea. A particularly compelling sort of experimental paradigm is one where people are asked to freely remember an event from their life, and then prompted to say *from what point of view* they are viewing the episode. In very many cases, a person's memory of an event includes seeing themselves from the third person—which, it goes without saying, cannot actually be what was perceived subjectively at the time. This sort of "perspective flipping" is documented in study by Nigro & Neisser (1983), who make a distinction between the "field" view (where the memory is described as from the first person field of view) and the

"observer" vantage point (third person). They note that "a deliberate attempt to remember the "objective circumstances" of an event leads to relatively more observer memories; a focus on feelings leads to more field memories" (1983: 481). Rice & Rubin (2009; 2011) have found that "observer" memories tend to be extremely common in day to day event memory (as opposed to recalling highly emotionally affective memories). They also remark that

> [a]n interesting pattern emerged when examining the predominant location used for each event. In several cases perspective location corresponded with the likely location of other individuals. Memories of running from a threat tended to come from behind the individual, whereas performing in front of others, either as a group or individually, produced memories from in front of the individual. (Rice & Rubin, 2011: 575)

Clearly, in these and similar cases, what is being "remembered" is not strictly the episode as experienced—the memory is a reconstruction of the event, based on the recorded perceptual stimuli at the time, but also modulated unconsciously by semantic information, and associations. A memory of a performance brings both the perceptual experience(s) encoded at the time, as well as semantic memory of what performances are ("something you watch" is probably the more available definition, rather than "something you do on stage", at least for non-professional performers).[16] Also, when we perform in front of an audience, at the time, we are probably partially (perhaps not consciously) imagining or simulating how we look *to the audience*, so that information may be part of the engram encoded at the time, and when we recall the event, the audience point of view may be more available, for this reason.

Connectionists, too, have to view memories as constructed in this sense. Ceci (1995) describes McClelland's connectionist approach, in which "a given memory is represented by a pattern of activation across neurons and connections, some of which are also part of the representations of other memories" (Ceci, 1995: 118). McClelland's (1995) *trace synthesis model* posits that traces are encoded in multiple nodes, but "bundles" can be created by associations and constellations of activation.

> The model illustrates two key points... First, it provides an explicit though simple mechanism illustrating how memory distortions can arise from the workings of ordinary memory retrieval processes. These processes are often beneficial—they allow the formation of generalizations over similar instances and the filling in of missing properties based on the properties of other, similar individuals—but they can be potentially harmful in that the information filled in need not be correct. Second, the model has the same property that human memory has, of often failing to separate information that arises from different sources. Suppose a new instance node is formed for every experience, and suppose one has a number of similar experiences.

> Then when we try to recall one, pieces of other similar experiences will tend to intrude particularly in those aspects of the original for which the information is weak or missing.... it is unfortunately not possible to inspect every memory trace individually; the information is not stored in the units themselves, but in their connections; like connections among neurons in the brain, we only know what is stored in them through the effects these connections have on the outcome of processing. But since many units and connections contribute to this outcome, full disentangling of the specific cause of each aspect of the outcome is impossible. It will, then, not be possible to identify the specific source of any aspect of constructed recollection. (McClelland 1995: 73)

"Disentangling" is a useful metaphor here. An act of retrieving *p* may well bring along a lot of things tangled up with *p*: various other "bits" of memory that are not the target of the current retrieval process, but, rather, traces of post-event information, or semantic connections to the retrieval context, or simply bits that sub-serve multiple memory traces, and hence can easily be activated, even sometimes by mistake. Note that this is reminiscent of Barsalou's account: certain obligatory associations embedded in cognitive frames can be activated even in contexts where they are not propitious, given that various aspects of a concept (or belief, or event) recorded in memory may be stored in several modes and locations, and those multiple modes may be blind to one another. There is no global, conscious cross-checking and error-correction program running to disentangle memories as they are retrieved. In short, nothing in this picture of memory storage and retrieval suggest a global central system of any kind. Rather, it sounds like exactly the sort of massively parallel associatively-driven subdoxastic functioning that one should expect from a modular encoding and retrieval architecture: it operates unconsciously (blindly) and almost exclusively via associative priming. Memory retrieval does *not* rely on objective searches, and global evaluative comparisons, but rather, retrieval cues impose contextual restraints to pull together various (associatively connected and compartmentalized) bits of data to reconstruct memory on-line. But then, how can *belief* fixation be a process of objective searching and global evaluative comparison if every such process relies, as a first step, on memory retrieval—recalling items from the cognitive background? Note that this question does not imply that beliefs are autobiographical memories of some sort: rather, my claim is simply that any act of belief fixation or revision will require the recalling of *something*—be it previous reasons for believing, previous experiential evidence, or merely stored semantic knowledge. In short, there is going to be a recall step that mediates *any* deliberative belief revision or

consideration process, and this recall step will *not* be characterized by objective searching and global evaluative comparisons: rather, it will be associative and highly contextually constrained.

It is for these reasons that I have argued in this section that Tulving's work on memory helps support the modular picture of revision I am endorsing: it dovetails nicely with the sorts of concept storage and associative connectivity between conceptual frames that make believing *possible*—i.e., allowing for tractable searches, consistency checking, etc. This is one of the main lines of argument I will pursue in the next section.

5.3 Modular belief

To recap briefly the arguments made so far in this chapter, my contention has been that belief formation is plausibly construed as "modular" in the sense that the process is made tractable by the modularity of the concepts which constitute beliefs: the idea is that if concepts are formed via modular processing, there will be an attendant organizational structure which allows for tractable search and association processes *between* concepts and which facilitates subsequent composition and revision. Fodor objects to this sort of argument, viewing any extension of the modularity thesis beyond the realm of the sensorium to be a mistake, as we have seen repeatedly throughout this and previous chapters. However, I have made the claim that Fodor's theory of concept acquisition via "attractor landscape" is much more clearly explanatory *only if* we invoke a modular architecture to subserve that system. Granted, Fodor would not agree to this claim, but for the moment I would like to simply assume that Fodor might provisionally grant this claim, and note that he will still have a fundamental objection to the *next* stage of my argument, which is to suggest that the acceptance of modular concept formation leads inevitably to the idea of modular belief formation.

5.3.1 Associative processing: how far can it take us?

Fodor will object that this second stage—the formation of belief out of constituent concepts—requires a global consistency check that simply cannot be effected by local (modular) computations. Modules are domain-specific and encapsulated, which renders consistency checking impossible: belief formation will require some sort of global, domain-general central system that modules are incapable of constituting. And it's not just

consistency checking that is at issue. Fodor will note that in order to take an attitude towards a particular concept that it *is true* (and hence to be believed), one has to examine some evidence. But, according to Fodor, the evidential domain is technically *the entire epistemic background,* which by definition, modular processors do not and cannot have access to. In short, trying to explain belief fixation via modular processes violates the isotropy of belief and the Quineanism of the process.

My response here would be to note that it is, of course, entirely correct that a module cannot have access to the entire epistemic background in the sense of being cognitively penetrable, or having the capacity to scroll through the concept store to examine relevant concepts in turn. However, a *suite* of massively parallel modules could have access to the entire background, in aggregation. As I have argued in §5.2, we can employ the ideas of "conceptual frames" (from Barsalou) and many of Tulving's insights regarding how concepts are encoded in and retrieved from memory to show how complex, composed conceptual structures (including elaborate episodic memories) are organized, tagged, and compartmentalized in a fashion that facilitates tractable searching and retrieval via associative priming. Furthermore, the processes via which this organization is managed appear to be modular, insofar as they are automatic, fast, unconscious, prone to systematic breakdown, etc. I contend that the "attractor landscape" idea also presumes a level of encapsulation and domain-specificity of function. Regardless, there is no reason to think a *belief* formation process would not benefit from exactly the same sorts of connections and compartmentalized filing that concepts and memory exhibit. Fodor's image of belief formation is one where (Quinean) central systems take a possible belief and proceed to weigh it against all relevant evidence and then cross-check for coherence with all other belief. I am suggesting that nothing of the sort needs to happen: rather, if a certain concept is associatively linked to other concepts that are already taken to be "true" (and hence believed), then it too will be believed, barring an association to some confuting evidence.[17] We do not need to check every belief against every other belief, only the relevant ones, and the relevant ones are those that are already largely compartmentalized and associated with the one in question.

Of course, a Fodorian will note the obvious problem in that schema: it's easy to suggest we only consider the "relevant" information, the problem is that determining what's

relevant in the first place requires a global view—which is precisely the point of the original *frame problem*. I would argue, however, that when it comes to evidence and coherence, what's "relevant" is baked into the organizational structure of the items that are already in the epistemic background—relevance *is* an association, after all—if something is relevant, in very many cases, an association will be encoded. Fodor admits as much in his account of concepts as files: files contain "memos" to other *relevant* files. The point is that you can make the connection more quickly between some concepts than others because of some shared aspect(s)—and crucially, these relevant associations are subdoxastically and automatically self-organized during concept formation, according to the attractor landscape account. The upshot of this is that we don't need to "determine" relevance between concepts and between beliefs (which would, indeed, seem like a Quinean process, and would have to treat belief as isotropic): rather, relevance is baked in from the start in many, probably most (and perhaps nearly all) cases.

Fodor's initial argument assumes that *all* instances of belief fixation are instances in which "*any* of one's cognitive commitments is relevant to the (dis)confirmation of any new belief" (2008: 115). I think that's a serious overstatement. I would suggest that in the vast majority of cases, the cognitive commitments relevant to any particular belief are already filed with (or associatively "close" to) that belief. This argument should be agreeable to Fodor: the constituents of belief are concepts, and the content of concepts (on Fodor's view) are properties in the world. If two or more concepts are *relevant* to one another, then this presupposes that there is a relevant connection between the properties (or what an agent can do with them) *in the world.* So, if two or more beliefs are relevant to one another, then it must be by virtue of the concepts that constitute those beliefs being mutually relevant. And this is a function of the world, and which is baked into the concepts via the attractor function of the acquisition stage, and preserved by the compartmentalized filing system under which they are subsequently named and organized. It is simply not the case that every belief is potentially relevant to every other belief, as Fodor demands. There is a hard limit to relevance between beliefs, set by the *world*—beliefs can only be mutually relevant insofar as their constituent concepts are in some way relevant, and those concepts can't be mutually relevant unless the properties in the world to which those concepts are locked are mutually relevant. In other words, only beliefs which are somehow conceptually connected (or

usefully connectable) will ever end up brought to bear on each other, and these will already likely be connected via conceptual associations (and inter-file memos) at many levels. The files, including the belief files, need not be isotropic, and arguably *aren't*, which explains why some files can be accessed more quickly and easily than others depending on context. If they were isotropic, that shouldn't be the case.

Fodor will argue that *some* belief contexts may be brand new in a way that connects certain concepts as relevant to one another in a *novel* way (thus requiring a form of non-demonstrative inference that can't be found by mere association). A Fodorian will note that, just above, I suggested that not *only* previously conceptually connected beliefs may be brought to bear on one another, but also connect*able* ones—this, the Fodorian might argue, implies an openness to the forging of novel connections, which means *in principle*, every belief could be brought to bear on every other after all (which is precisely what Fodor said in the first place). However, I would defend my point by noting that connect*ability* implies some level of pre-existing connection (or the at least the preconditions which allow for connection). Even if that connection has not been made conceptually, one must assume that the connection can be made (in some sense) *in the world* (i.e., via some application). Every "Eureka!" moment of novel scientific discovery is a new *conceptual* connection, but it merely *discovers* an existing connection in the world. And since concepts get their content from locking to properties in the world, then "connectable" concepts are already technically connected, even if they haven't been *filed* or *recognized* as connected (yet).

The same principle should apply to all the sorts of reasoning Fodor declares must be Quinean and isotropic: abductive inference, analogical reasoning, etc. Consider Fodor's claim regarding analogical reasoning:

> "analogical reasoning" would seem to be isotropy in the purest form: a process which depends precisely upon the transfer of information among cognitive domains previously assumed to be mutually irrelevant. By definition, encapsulated systems do not reason analogically. (Fodor 1983: 107)

Yes, the domains in questions were previously *assumed* to be irrelevant. Yet, if the analogical reasoning is successful, then apparently the two domains (in fact) were not mutually irrelevant. My argument is that the concepts stored in those cognitive domains had within their "files" the information necessary to connect them—it might just take some time for the perfect constellation of associative strength to bubble up when the two are co-

tokened. The associations are there, and likely activated, whether they are *noticed* or not. The fact that there is a point in time where they *aren't* noticed just adds further support to the idea that associative processing must be pretty circumscribed and limited in terms of how many associations can be brought together in consciousness at the same time. And whatever process is managing the ascendance of some associations (rather than others) apparently isn't very "smart"—because it fails to notice the connection, perhaps indefinitely. Note that often, after the connection has been made, we are struck with the sense of "how could I not have seen it earlier?" The sense that the connection (or analogy) was apparent from the start, yet somehow missed, can be overwhelming. The upshot of all this is that determinations of relevance can be built in increments: an even very minor degree of (automatic) association can be bootstrapped up into a much stronger degree of relevance if provisional deployment of the association proves profitable.[18]

5.3.2 Belief revision and the limits of recall

We can provide further support for this idea by digging a little deeper into Tulving's account of memory retrieval. One major *constraint* on retrieval is what Tulving & Thomson (1973) originally called the *encoding specificity principle* which "...emphasizes the importance of encoding events at the time of input as the primary determinant in the storage format and retrievability of information in the episodic memory system" (Tulving, 1972: 392). Recall that the process of retrieval is predicated on the interaction between the retrieval cue and the information stored in the engram—and this interaction will be inhibited or amplified based on the "compatibility" of the cue and the engram: the more compatible the cue and the engram are, the more available that engram is for retrieval. Now, as discussed above, in the process of retrieval, the cue will be affected (unconsciously and subdoxastically) by other *semantic* associations activated at the time of retrieval. Similarly, the initial encoding of the engram takes place within a context of certain semantic associations as well. Tulving (1972: 224) explains that "the engram of a stored event in the episodic system, and the retrieval cue, as interpreted or encoded in light of the information in the semantic system, must be compatible for remembering to occur." The upshot of this is that we cannot recall things unless the context of retrieval is in some way matched to the context of initial encoding, which means that we can't recall something without already having a relevant connection in

mind. Irrelevant items simply *are not automatically retrievable*. This, immediately, has a tractability payoff in searching memory for relevant beliefs. Fodor insists that every belief is in principle relevant to the (dis)confirmation of every other belief and that we'll therefore need an unencapsulated system to check across unassociated cognitive domains, but the reality is that if this checking procedure utilizes memory retrieval (which clearly it must), then that process will be circumscribed by the limitations the memory system places on retrievability, and the associative priming on which it runs. Encoding specificity imposes domain specificity on memory retrieval in a way that cuts against the global, domain-generality of Fodor's putatively non-modular central systems. But once domain-generality in belief revision is given up, then one might as well stop resisting the modular argument that fits so well with everything else under discussion in this chapter.

Numerous studies have backed up Tulving's *encoding specificity principle*. Godden & Baddeley (1975) present clear evidence that recall is better when contexts of encoding and retrieval are matched. They had deep sea divers learn and memorize word lists both on land and underwater—the divers' recall was improved when tested in the congruent environment (i.e., words learned underwater were recalled better when underwater than when on land, and vice versa). This, of course, is probably not surprising to most of us: many people use spatial/environmental cues to aid recall. When I can't remember where I put my sunglasses, I *retrace my steps*—of course, doing so physically I may just stumble across them; but it also often happens that merely getting *close* to where I left them, I suddenly remember. Eich & Metcalfe (1989) show that the context of one's *mood* matters as well—both in the sense of mood to subject matter congruence being important (i.e., the material being studied matching the mood of the learner at the time of encoding) and diachronic mood congruence (i.e., the person being in the same mood when tested as when encoding took place).[19]

Hannon & Craik (2001) stress the importance of semantic congruence between encoding and retrieval. Whatever semantic associations were activated at the time of encoding the information originally will need to be *matched* in the retrieval cue for effective recall. For example,[20] if you have subjects remember a list of words including ACCOUNT TELLER BANK, when you prompt them to produce the remembered words later, recall of BANK will be much improved if you prompt them with MONEY rather than RIVER. This is because the second definition was not semantically activated at encoding. Furthermore,

Craik & Tulving (1975) show that *elaborative* semantic connections can take place at encoding, which can have major effects on subsequent retrieval. For example, learning word pairs that are connected in a sensible way makes them more easily remembered (i.e., "furniture-settee" is more easily retrieved than "jungle-potato"). The idea is that the "more richly semantically elaborated" the trace is during encoding, the richer the field of associative cues that may match to it during retrieval (Hannon & Craik, 2001: 240). Elaborated cues, e.g., items that fits into causal arrangements, or fall under similar conceptual frames, or fit narratives already established—that cohere, essentially—will be easier to recall because its more likely that that, being coherent, the context of the retrieval cue will come closer to matching the encoding context.

So the *encoding specificity principle* can help explain successful and unsuccessful recall in a way that fits well with the account of compartmentalized storage and priming via association that I am defending. Coherence aids recall (and non-coherence inhibits recall), which goes a long way toward explaining how a *coherence checking* procedure can gain tractability, automatically and subdoxastically. It isn't Quinean after all: by merely evaluating a belief for coherence, I will need to recall my belief (and presumably the circumstances of, or reasons for, my coming to believe it). But as we have seen, this recall context will automatically prime *other memories* (including other beliefs, and the reasons for those, and so on) that *already cohere.* What this suggests is that if I evaluate a belief for consistency, I will often very quickly find before my mind the very beliefs I most need to check it against. And this happens automatically and with no central, global executive. If, on the other hand, there is no connection between retrieval context and encoding context – *nothing gets remembered.* But this means that beliefs can't be isotropic in the way Fodor suggests, assuming that beliefs are stored in memory. The belief you are trying to check will affect the retrievability of your other beliefs, but you won't have any degree of control over that effect. So it is possibly (and presumably often) the case that some beliefs may not be able to be brought to bear on some others, as the retrieval of one inhibits the retrieval of the other from memory. So our belief revision practices cannot be Quinean in the full sense, even if they ought to be.

5.3.3 Systematic patterns of breakdown

Let me try and bring together the threads of my argument here briefly. I am suggesting that the constraints imposed on conceptual filing and memory encoding (constraints that have all the hallmarks of being the work of massively parallel, modular processes) make it such that tractable search and retrieval functions can subsequently take place. A modularly-circumscribed concept acquisition process results in concepts that are stored and filed in self-organizing structures with multiple interconnections and associations built in as a feature of the way they were encoded and stored. This account seems to be supported by Tulving's work on memory encoding and retrieval, as well: insofar as belief revision involves *recalling* previous cognitive commitments, the nature of memory retrieval is such that what's available for comparison is automatically and subdoxastically limited by the context of the recall cue. In other words, memory appears to be anisotropic and recall is structurally non-Quinean. Merely *considering* a belief will automatically activate the stored constituent concepts, and the retrieval of those concepts from memory will be shaped by the retrieval context (the "cue") in a way that essentially *brings to mind* the relevant associated concepts automatically and in a volume small enough that it can be tractably sifted. However, the price of this tractability is that we need to give up on the Fodorian insistence that belief be construed as isotropic and Quinean. No belief is, in actuality, globally available to *every* other belief, as there is a recall step that mediates that availability, and the recall step runs on associative processing—it's not Quinean. But, as I have argued since chapter 2 of this dissertation, no computational system can achieve the sort of Quineanism Fodor wants, anyhow: belief revision in order to be tractable has to be heavily circumscribed, and this will result in normatively sub-optimal belief fixation and revision.

In short, in order to succeed *at all* in belief revision, we need a system that is also prone to certain failures. These are the predictable failures of a system that runs on (incomplete) associations. We make do (we satisfice), and it doesn't always work. We have examined Fodor's argument that belief formation requires global access, and modules can't *get* global access, hence belief formation can't be the result of local, modular processes. My response is that the way out of this inconsistent triad is not to drop the modules, but to deflate what we mean by "global access". No module has global access. But all the modules, taken together, have global access, of a technical sort. And if they work in parallel, then we can get

an approximation of global searching and evaluation, yet in a way that tractably runs on local computations. It will not be perfect or optimal, and where it fails, it will leave us with self-reinforcing (coherent) webs of faulty belief (which happens all the time), or unawareness of an unchecked inconsistency (which happens all the time), or simple blindness to a connection that is just waiting to made that would (dis)confirm the belief in question if only it had been activated (which happens all the time).

5.4 Review and look ahead

Belief in the *normative*, Quinean sense *should* be isotropic. But this is a reason to let go of the normative picture, as it seems clear that in practice, belief is simply not isotropic. Belief is constituted by concepts, and managed by memory in ways that make it *not the case* that any belief can (in principle) be brought to bear on any other belief. Only associatively linked beliefs can be brought to bear on one another. And given the way concepts (which are constitutive of belief) and events (of which previous acts of believing are a subset) are stored and retrieved from memory, associative links are baked in to belief, and hence make belief management tractable. The normative Quinean picture of belief revision can still serve as a valuable regulative ideal, but it is one that a human mind will only ever be able to approximate. And the best account we have on hand to explain *how* a cognitive system can approximate global, Quinean belief revision without actually being global or Quinean is to invoke modularity through and through.

I have attempted to show in this chapter that even Fodor's own account of concept acquisition demands a massively modular architecture to explain how concepts are locked in place and organized in storage. I have *not* made an argument that there is a single "concept acquisition module"—to do so would be explanatorily unhelpful. Rather, what I am pointing out is that these processes—which Fodor would be the first in line to claim *cannot* be modular in nature, due to their seemingly Quinean, isotropic nature, and global access demands to the epistemic background—certainly exhibit "modularesque" qualities and characteristics and patterns. And the reason for that, I believe, is that these processes of concept acquisition and revision are entirely subserved and mediated by lower level modular processes and more specialized modular integrative functions: inheriting all the benefits of the informational encapsulation, domain-specificity, and fast automated processing that this

entails. This, I argue, is the only way to a) explain the observable data, and b) explain *in principle* how what we observe could take place without running square into the frame problem and violating the demands of computationally tractability. Given both the *finitary predicament* and the *facts of what we achieve in cognition*, the explanation that all of it is the result of associative processing among massively parallel lower-level modular mechanisms, including multiple layers of modular integration and interface functions, is the only explanation on the market that gets us off of the horns of the dilemma.

However, this explanation still leaves a few details left to be better explained, rather than simply waved away as "the work of modules" or chalked up to associative priming. Specifically, everything discussed in this chapter regarding associative processing assumes that associations can co-activate items (concepts, memories, beliefs) that are mutually relevant for further processing—which implies there is some sort of *space* within which these items are brought together. A large part of this chapter has been devoted to explaining tractable search procedures during belief fixation, though we have yet to explain the *destination* of the items retrieved from these searches. Additionally, there is much more that needs to be said about the specifics of the sorts of search algorithms that can approximate global scope while maintaining frugality with respect to limited cognitive resources.

In the next chapter, I will set out to build my own positive case for an architectural framework that I believe could, in principle, provide the 'globality on a budget' that we are looking for, and which has a high degree of intuitive appeal and evidentiary support. This case will involve two key elements: a global workspace to which disparate items from the cognitive economy may be brought for processing, and a nested set of heuristic search algorithms to manage entry into that workspace. The combination of these two features forms an account that can get us as close as possible to an approximation of the Quinean ideal of bringing all and only the relevant information to bear on any given deliberative question.

Notes for chapter 5

[1] Fodor credits Chomsky for being the first to hint at this problem, insofar as Chomsky recognized that *associative* mental properties (which by definition work with representations *across* domains, i.e., not in direct linear causal arrangements) seem impossible in a mind that has *productive* mental processes. Chomsky's critique of Skinner was based in part on the observation that "the mind is sensitive to relations among interdependent elements of mental or linguistic representations that may be arbitrarily far apart. Since association is *contiguity*-sensitive, such relations can't be associative" (Fodor 2008: 103).

[2] Carruthers (2006a) calls this "wide-scope encapsulation" as modules are, after all, "sharing" information. The point is that the system viewed in its totality still is as heavily informationally circumscribed as the modules underlying it are, in aggregation.

[3] One can trace hints of it as far back as Aristotle, and explicitly find reliance on a LOT hypothesis in the work of many philosophers commenting on Aristotle in the Medieval period (Buridan and Ockham both being obvious references. Indeed, much of Buridan's theory of mind dovetails surprisingly well with Fodor—Buridan's 700 year old account of how concepts are abstracted from "vague singular impressions" in his *Questions on Aristotle's de Anima* (*Book III*) sound remarkably similar to the story of concept acquisition that Fodor spins in *LOT 2,* and which is the subject of this section.

[4] I will follow Fodor's lead here and *not* spend any time defending RTM here. As far as Fodor is concerned, "RTM remains the only game in town" (Fodor 1998: 23), and has been central to the account of all "mental realists "arguably since Plato and Aristotle, patently since Descartes and the British empiricists" (Fodor 2008: 6). Since this section deals with Fodor's views, and my goal is not necessarily to defend those views, but merely to show how Fodor's account answers it's own objections, I will assume what he assumes without discussion in order to get that argument off the ground.

[5] Fodor's "five not-negotiable conditions on a theory of concepts": concepts must be 1) causally efficacious mental particulars, 2) able to be employed as categories, 3) compositional, 4) sometimes learnable, 5) public (1998: 23-28).

[6] Note that this "doorknob/DOORKNOB" problem is a version of what is referred to in the psychology literature on memory under the term the *binding problem*—which refers to the problem regarding how separately encoded features of particular memories are bound together in order to create single memories of particular events. As Zimmer *et al.* (2006: 3) suggest, "the binding problem is a ubiquitous one that has to be solved in perception and in action; it is also a problem in memory because binding of features is necessary during encoding, consolidation and retrieval... the temporal synchronization of the discharges of individual and feature-specific neurons which form dynamic cell assemblies." We will look at this in some more detail below, when the discussion turns to storage and retrieval of concepts (and by extension, beliefs) in memory. Stainton & Viger (2000: 141), in their review of Fodor's *Concepts,* similarly point out the connection to the issue of binding.

[7] Compositionality is required if we hope to explain the productivity of language and thought—since productivity requires the ability to combine concepts to form ever more complex concepts and propositions. The bottom line is that compositionality will be the litmus test for any theory of concepts "since mental representation and linguistic meaning are *de facto* compositional, we can reject out of hand any theory that says that concepts (/word meanings) are Xs unless Xs the sorts of things for which compositionality holds" (Fodor & Lepore 2002: 3). Fodor argues in *LOT 2* (as he does repeatedly elsewhere) that the only theories that can meet the challenge of compositionality require the idea of *conceptual atomism* (for example, you can't acquire the concept BROWN COW without the concepts BROWN and COW to compose it out of. But you will have to take BROWN and COW as primitive, or atomic, or else you will never get off the ground. Even if you think BROWN and COW *could* be decomposed somehow (perhaps via definition), then you are just pushing back the point where you have to accept *some* concepts as atoms (the ones that form the constituent parts of BROWN and COW) in order to start concept-building).

[8] Stainton & Viger (2000) similarly note that Fodor's explanation of this process seems incomplete insofar as "Fodor concentrates so much attention on the doorknob/DOORKNOB constraint that he seems to forget to fill in the rest of the story of concept acquisition" (2000: 144). They also note that his insistence that concepts can't turn out to be stereotypes seems to incorrectly capture his own discussion of concept acquisition. Note that

below we will see how Fodor arguably remedies some of these issues in his (2008) *LOT2* revisitation of the account.

[9] Fodor notes, with some irony, that he is stealing the idea of an 'attractor landscape' from the connectionist literature (which he is generally dismissive of). For elaboration on what exactly an attractor landscape is, there are good discussion in Christiansen & Chater's *Connectionist Psycholinguistics* (2001), and Jeff Elman *et al.'s Rethinking Innateness* (2001).

[10] One question we might ask here is *what happened to the original DOG concept—the one that overgeneralized?* It's not clear on Fodor's account what we are to make of that question: the quick answer would be that the child (who still mistook lambs for dogs) simply did not have the concept DOG yet. But she did have *something*—namely, whatever concept she subsumed dogs and lambs under (FOUR-LEGGED FURRY CREATURE?). If this is a concept, then it needs it's own whirlpool. Or else it's not a concept, but rather some form of proto-concept that gets elaborated *into* a concept later on, or simply serves its purpose as a temporary bridge and then is discarded once valid concepts are formed. We will return to this question below when I discuss the issue of belief *replacement* vs. *revision.*

[11] Here's where Fodor can interject and say that I just invoked a global sounding operation: if concept acquisition is modular, as I am arguing, then we *should* see cases of individual concept breakdown, since there should be a dedicated piece of neural real estate for each concept-acquiring module. To this I would respond that what it shows is, rather, that the modularity of the concept acquisition system is massively parallel.

[12] Just to make it clear why this process makes sense in terms of balancing optimization and usefulness: it's *useful* for a child to learn a lot of individual words, long before figuring out the general rule which applies. But this means, in the short term, each word will essentially be its own conceptual file—in its own "compartment". Eventually, as vocabulary expands, this will be intractable, as there are too many isolated compartments to search through to find the words one needs (potentially thousands). So words are subsumed under general rules—like how to generate past tense constructions. Now the storage problem is solved, but a *new* problem is introduced: the irregular forms have been switched to the regular form and are now *wrong*. Slowly, these few dozen verbs need to be refiled back in separate compartments (given their individual constructions). This seems a pretty good trade-off that keeps storage tractable, yet still allows for *use* at early stages.

[13] Tulving's distinction between *episodic* memory and *semantic* memory, it is probably pretty obvious enough from the names. Examples of episodic memory would be things like my remembering having gone snowboarding at Mt. Tremblant last February, or remembering that I need to return the library books on my desk before next Friday, or remembering the taste of the chocolate cake I had after dinner. All of these memories involve temporal-spatial relations, and they all involve *me* and my experience in some sense. Examples of semantic memory would be such things as remembering Mt. Tremblant is just over an hour's drive North from Montreal, or remembering that next Friday comes before next Saturday, or remembering that cakes are baked in ovens, and don't grow on trees. These latter semantic memories may *involve* temporal-spatial relations, but they involve them on a *cognitive* level, and don't require perceptual experience to constitute them. This is clear from the fact that I can come to possess the semantic memories above *without* having the episodic ones (i.e., I can simply *know* those things). An individual event/experience can encode differently in the two systems: for example, Zimmer *et al.,* (2006: 11) cite McClelland & Rumelheart's (1985) study which demonstrated that subjects shown a blue banana will easily form an episodic memory of the event, but the *semantic* memory—knowledge—of bananas does not update accordingly (i.e., it does not accommodate the event and "learn" anything new about bananas).

[14] Interestingly, Tulving argues that the metaphysical specifics of what exactly engrams *are* is fairly irrelevant as far as his account goes—one can remain relatively agnostic on the question—though he says that the position one commits to on this question *will* have an effect on the sorts of questions one seeks to answer:

> Whether we think of engrams as information stored about past events, as a record of operations, attunements, or dispositions, or even as pictures, images, copies, propositions, analogue representations, feature bundles, or as particularly marked parts of associative networks, makes relatively little difference to our understanding of how memory works, although it may influence the thinking of individual students of memory, the kinds of questions they pose and the kinds of data they collect. (1983: 160)

Nevertheless, Tulving admits he is "partial to the idea that the engram of an event is a bundle of features."

[15] The idea that memory is constructed rather than merely retrieved goes back to Bartlett (1932); Tulving is not the first to suggest it.

[16] Actually, I would speculate that when professional performers recall performances, they might be less likely than non-regular performers to perspective-flip. I would recommend an experiment to test that hypothesis.

[17] Recall Gilbert's 'Spinozan' account of mind: that we must first believe a proposition before evaluating it (as discussed in chapter 1). Perhaps this is a simpler, more elegant way to make the same point I am making here. I am quite sympathetic to Gilbert's account.

[18] Note also how many advancements in human thought are often independently "discovered" or worked out in roughly the same time period. This seems a similar process of relevance-creep by degree: there comes a point where two people, working independently, but from roughly the same epistemic background, put things together in the same ostensibly unanticipated, novel fashion.

[19] Emotional congruence can also improve reaction times for associative prompts (Spezio & Adolphs, 2009: 91). *Cf.* Niedenthal *et al.* (2002) and Ryan & Eich (2000) for further examples.

[20] This is my example, not Hannon & Craik's.

6 Globality on a budget

I have been trying to trace out an account of concept (and by extension, belief) acquisition that maintains computationally tractability via modular integration, and specifically employs Fodor's theory of concepts to show a way around his concerns about the frame problem. Of course, Fodor argues that anyone who thinks they've gotten around the frame problem has likely just begged it in another way:

> I do seem to be going on about this. That's because it strikes me as remarkable, and more than a bit depressing, how regularly what gets offered as a solution of the frame problem proves to be just one of its formulations. The rule of thumb for reading the literature is: If someone thinks that he has solved the frame problem, he doesn't understand it; and if someone thinks that he does understand the frame problem, he doesn't; and if someone thinks that he doesn't understand the frame problem, he's right. But it does seem clear that whatever the solution of the frame problem turns out to be (if it is just one problem; and if it has a solution), it's not going to be computationally local. You usually can't tell from the local (e.g. compositional) structure of a thought what is relevant to its (dis)confirmation. Clearly, you have to look at a lot else too; the frame problem is how you tell which else you have to look at. I wish I knew. (Fodor 2008: 120-121)

Fodor actually isn't wrong in the first half of this quote—the frame problem really is a sort of zombie problem that keeps rising from every attempt to beat it back. We will even see in this chapter, despite further attempts to avoid it, there will be small cracks in the frame each time. Yet I still disagree with Fodor on the second half of that quote—I will still be arguing that there is a way out of the frame problem, and a way to explain how we revise belief, engage in abductive inference, think creatively, all of it, constrained by entirely local computation. It will mean giving up something up: optimality. And it will involve a deflationary understanding of what "global" processes really are (hint: not Quinean, but they can mimic Quineanism). We will have to settle for "globality" on a budget—that will have to satisfice.

In the sections that follow, I will attempt to sketch a picture of how cognition might be structured in a way that *approximates* the norms of belief revision and rational thought, while still maintaining plausible computational tractability and skirting the *frame problem.* The account requires answering two questions:

- What sort of *global workspace* is available for bringing disparate systems into contact with one another?
- What sort of heuristic algorithms can expedite and/or limit searches, and tractably manage the global workspace?

6.1 The global workspace

Recall Cherniak's argument that "only beliefs in short-term memory can be premises in reasoning; beliefs in long-term memory are inert" (1986: 59). This admonition, if it's correct, applies to *all* reasoning—both the unconscious, automated, subdoxastic work of System 1, and the controlled, conscious, rule-governed reasoning of System 2. So what is the nature of this short-term memory space—or following Baddeley (1986), what we might better refer to as *working memory* space—into with certain beliefs, memories, propositions are pulled to be engaged in reasoning processes? In the previous chapter's discussion of Tulving's theory of memory, we saw that memories are reconstructed via the interaction of a retrieval cue with stored episodic and semantic information (in the engram); but we might ask *where* this reconstruction takes place. The standard answer to this question invokes some sort of "global workspace"—either metaphorically, or as a functional description, or as a discrete physical space in the brain where representations converge for processing. For our purposes, I will discuss the global workspace hypotheses that we will look at in this section at the level of functional description, and make no particular claim regarding how or where they are realized.[1]

6.1.1 Blackboard Architecture

The idea of a "global workspace" is often credited to Baars (1988; 1997), who envisions a "space" mediated by working memory which is "global" only in the sense that anything can, in principle, be brought there. It's a *passive* space, not an active central system that seeks out anything. Here is Baars' basic description of the workspace:

> 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).

The vision here is of a "blackboard architecture," where one expert, working independently, might have a result that another expert can usefully employ (and wouldn't have "thought" to use otherwise). In fact, it's quite possible that one expert could report findings that *don't even make sense* or serve any use to that expert, but are similar enough, or associatively linked to another expert's domain, such that the other expert picks it up. In this way, the system as a whole might get useful information *for free*. One might call the effects of such a system a sort of *cognitive spandrel*, echoing Gould & Lewontin (1979)—the idea would be that one system processes perceptual information according to its own program, and outputs the result, including elements that are not necessarily relevant to that process, nor to the next stage of linear processing that the system is designed for. However, that "irrelevant" information, once in the global workspace, might be picked up by another system that *can do something useful with it*, yet would otherwise never have had it as input.

Many theorists have embraced Baars' blackboard architecture in an attempt to explain how various conceptual integration and composition functions might take place. Pylyshyn (1999) uses it as a valuable way to model the language and visual systems' coordination of separately processed and interpreted features, noting that the employment of a blackboard architecture has been very successful in designing *machine* systems for vision and speech recognition:

> [These] systems use a so-called "blackboard architecture" in which a common working memory is shared by a number of "expert" processes, each of which contributes a certain kind of knowledge to the perceptual analysis. Each knowledge source contributes "hypotheses" as to the correct identification of the speech signal, based on its area of expertise. Thus, for example, the acoustical expert, the phonetic expert, the syntactic expert, the semantic expert (which knows about the subject matter of the speech), and the pragmatic expert (which knows about discourse conventions) each propose the most likely interpretation of a certain fragment of the input signal. The final analysis is a matter of negotiation among these experts. What is important here is the assumption that the architecture permits any relevant source of knowledge to contribute to the recognition process at every stage. (Pylyshyn, 1999: 10)

There is some empirical evidence to support the existence of a global workspace in the human brain, mediated by consciousness. The standard experimental design that is employed

to try and isolate the working of a global workspace is a "contrastive analysis" (Baars, 1988) between conscious and unconscious states,

> wherein closely matched conscious and unconscious conditions are compared in waking subjects, either by stimulus manipulation (binocular rivalry, masking, attentional tasks, etc.) or by overpractice of automatic habits. In all cases tested so far, the conscious condition recruits very widespread cortical resources while the matched unconscious event typically activates local regions only... In a complementary experimental paradigm, brain response to stimulation has been compared in conscious versus unconscious states. Unconscious states studied include sleep, general anesthesia, epileptic loss of consciousness and vegetative states. Sensory stimulation in all four unconscious states evokes only local cortical responses, but not the global recruitment characteristic of sensory input in conscious subjects (Baars et al., 2003). This general pattern of results has now been shown for vision, hearing, pain perception, touch, and sensorimotor tasks. It appears that conscious events recruit global activity in the cerebral cortex, as predicted by the theory. (Shanahan & Baars, 2005: 166)

Of course, this evidence is not dispositive, but I think the idea of a global workspace or blackboard architecture is immensely helpful in reverse-engineering our own reasoning tasks. The success of artificial vision and speech recognition systems designed using these principles is suggestive of why it is an apt analogue to human cognitive processing—note how Pylyshyn (above) mentions how "the architecture permits any relevant source of knowledge to contribute to the recognition process at any stage". This blackboard architecture allows for exactly the sort of processes that approximate the Quinean, isotropic ones that Fodor demands. His insistence that all relevant background knowledge may be brought to bear on the process runs squarely into the frame problem, and yet in artificial systems, if they are set up as a series of dedicated modules ("experts") which feed into a global workspace, it seems to be working. Computer vision and speech recognition systems *aren't* getting bogged down by "terminal abduction", to borrow Fodor's expression.

Of course, speech recognition and vision might not seem like *reasoning* tasks, and fair enough. How does a global workspace design help explain how global reasoning tasks can take place—what are the "experts" who submit proposals and "negotiate" in the workspace in the case of abductive inference, or belief fixation, for example? This seems entirely more complicated than matters of sensory integration. Furthermore, sensory integration comes pre-framed via the domain-specificity of the sensory modalities, of which there is a finite, manageable number. But if we want to utilize the global workspace to manage *beliefs*, we seem to be exponentially expanding the number of access points we are

expecting to make use of the space. A "roomful" of experts with a blackboard is one thing—a *world full* of experts is another thing entirely. That's going to be a big, unmanageable blackboard—one that we likely won't be able to build a frame around.

Baars' original account of the global workspace suggests that the workspace is "managed" to some extent by consciousness. As Gilchrist & Cowan (2010) describe it, "according to this model, there is bidirectional information flow, with conscious processes influencing unconscious processes, and vice versa" (2010: 23). Conscious awareness for Baars, serves as a "spotlight" in the "theater" of the global workspace.

> It seems that the single most prominent function of consciousness is to increase access between otherwise separate sources of information... Everything is connected to everything else, via the bright spot onstage; that may seem to be a problem because it threatens to reduce our carefully evolved framework into an undifferentiated theoretical soup. But each element can be defined operationally, in terms of distinct observable events. These elements are usually separate from each other. Further, many elements such as "self," "working memory," and "sensory input" also have routine inter actions that are unconscious and therefore quite fixed. While everything *can* interact with everything else, it cannot do so with infinite flexibility. To allow such universal access we need that little bright spot onstage. But the bright spot of consciousness does not have very fast throughput. It creates a bottleneck that slows thing down. It is, in turn, influenced by the attentional network and other mechanisms. (Baars, 1997: 163-64)

I think this actually doesn't help us much to figure out the framing issues, however: it really does make it too easy to have everything connected to everything else through "consciousness". A Fodorian will object that this merely begs the frame problem, and that objection would seem merited in this case. A "bidirectional" workflow runs counter to the sort of *in*accessibility relations that must exist between modular systems—the domain specificity and informational encapsulation that grant tractability. The sort of "consciousness" Baars invokes here seems dangerously homuncular.[2]

Shanahan & Baars (2005) have a slightly different, updated description of the global workspace, which dispenses with talking of "consciousness" and tries to pin the frame issues back down to a modular substrate. In the (2005) version, Shanahan & Baars attempt specifically to explain how contextually relevant information finds its way into the global workspace, responding directly to Fodorian concerns that there is a sort of "input problem" lurking in that process.

> Fodor says little about the computational model behind his claim that informationally unencapsulated cognitive processes are computationally infeasible. Yet there are strong hints of a commitment to a centralized, serial process that somehow has all the

> requisite information at its disposal, and then has the responsibility of choosing what information to access and when to access it… By contrast, global workspace theory posits multiple, parallel processes that all contribute actively to cognition. Consider the computational processes that might underlie the likening of a Rorschach inkblot to, say, an elephant. Fodor's argument hints at a centralized process that poses a series of questions one-at-a-time—is it a face? is it a butterfly? is it a vulva? and so on—until it finally arrives at the idea of an elephant. Instead, the global workspace model posits a specialist, parallel process that is always on the lookout for elephantine shapes. This process is aroused by the presence of an inkblot that actually resembles an elephant, and it responds by announcing its findings. The urgency with which this process commends itself means that the information it has to offer makes its way into the global workspace, and is thereby broadcast back to all the other specialist processes. (Shanahan & Baars, 2005: 168)

I think this "inkblot" story is a lot better than the spotlight of consciousness version in terms of describing a plausible sounding process. However, there are still some hand-wavy elements here—specifically the point that parallel specialist processes are "always on the lookout for" whatever input arouses them, and that results from the workspace are broadcast back, globally. It seems a strange sort of architecture that would have every possible process running on active stand-by, just waiting for an invitation to enter the workspace, and constantly monitoring the global broadcast, like vigilantes listening to a police scanner. But the alternative, where processes are inactive until called upon, gets us back to the Fodorian picture, which is unframed and intractable (i.e., what does the "calling"? How does it know who to call? What is managing this mental rolodex?). Perhaps a better metaphor would be the proposal of Barrett (2005; also highlighted by and employed in the account of Carruthers, 2006a)—the "enzymatic" analogy. Barrett suggests that there is no need for active sorting of stimuli to appropriate modular inputs. Rather, modules can passively "find" the stimuli that activate them in a fashion analogous to the process via which enzymes build proteins within cells. Note that Barrett & Kurzban discuss this strategy in terms of inputs finding their way to appropriate modules (in response to Fodor's input problem, as discussed in chapter 4), though the same sort of enzymatic analogy could certainly be helpful regarding the current question: how does all and only appropriate representational content get into the workspace at contextually relevant and useful times?

> Enzymatic systems in biochemistry suggest an analogy with cognitive modules. Enzymes with diverse functions and diverse processing criteria can have access to a single common pool of substrates, or 'inputs', and yet still achieve specialized processing. Each enzyme has a recognition site that is capable of selecting its own inputs, or substrates, via a 'lock and key' template matching system. This means each device is sensitive only to its proper inputs and therefore can select its own inputs

> form a common pool. *No 'metamodule', or routing system, in necessary in principle* (Barrett & Kurzban, 2006: 634, emph. added).

This is perhaps more promising that Shanahan & Baars' description, in that it posits a seemingly much more domain-specific and encapsulated way to get the appropriate inputs into the workspace for a given processing context.

6.1.2 Semantic promiscuity

Jackendoff (2002) offers his own version of global workspace theory that focuses on associative and semantic priming to explain how various propositions or concepts generated via "parallel architecture" end up on what he calls the "global workbench of working memory" for integrative processing. In *Language, Consciousness, Culture,* (2007) Jackendoff explains how this works with regard to linguistic meaning construction, subserved by modular lexical parsing and production systems:

> 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 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).[3]

Memory priming is a key element of this: this 'parallel' processing architecture facilitates memory searches, and fits with the constructivist, associative processing described in Tulving's account. In essence, on Jackendoff's account, any given concept will 'light up'—associatively recall—any other memories that may be related, however tangentially, *regardless* of the current context. All of these memories would then be made available to appropriate interface and integrative modules which could run them through, strengthening some structures which seem more stable given the ongoing flow of context or communication, and inhibiting dead ends. And this process can be recursive: each time the strongest associations can be added to the recall cue, to re-run the search and activate further associations, promoting the stronger and inhibiting those that fall beneath some parametrizable activation threshold. 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).[4] 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 context awareness and relevance determination function). Rather, the promiscuity theory holds that numerous representations are constantly being cycled up from long-term memory to working memory on the basis of brute semantic or syntactic associations, no matter how strong or weak, and *regardless* of context (often in "obligatory activations" as discussed with reference to Barsalou's conceptual frames in the previous chapter). Recursive iterations can whittle the context down as the process goes on, but only as a result of cycling primed representations through various interface and integrative modules and, roughly speaking, seeing what sticks (or what achieves "resonance" in Jackendoff's terms). 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 domain specific, encapsulated processors to *recognize* inputs appropriate to them (every perceptual module does that, at the very least). The unframed version of the story is that either *(a)* some (global) process sorts and assigns inputs to appropriate modules, or *(b)* the modules search (globally) for the inputs that activate them. But, rather than viewing it in terms of *(a)* or *(b),* the account here, following Jackendoff, is that brute associative processing and semantic priming cycles up a "shortlist" of inputs that *might* be relevant to the processing task at hand. From there, dedicated modules, if they "see" their input in the global workspace, will be activated. The relevance of input to module is in this case tagged by association, not a deliberative process—hence there is no danger of slipping in what sounds like a global executive function, or "sorting box" as Fodor complained in the "input problem" discussed in chapters 4 and 5.[5] In some cases, modules will be activated by what turn out to be irrelevant associations, but if those modules, in turn, are organized in multiple levels of assemblies with other modules, then *relevant* processing will "resonate" in the sense

that certain assemblies will roar to life, as *all* the appropriate subcomponent modules are being activated. Irrelevant activations will run into a dead-end, as integration and interface devices higher up the chain don't take them up, lacking some complementary, co-activating input. Any simple example of homonymy, such as processing the hearing the word *bug*, as discussed above, demonstrate how this process plausibly proceeds. 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).

6.1.3 Taking stock of the global workspace

Let's quickly review: the global workspace account helps to explain the *how* and the *where* of tractable processing that achieves quasi-global scope without entailing truly global search and sorting algorithms. Such a workspace is necessary for any process of belief fixation and revision, such as I am defending, that is entirely subserved and mediated by a suite of massively parallel integration devices, designed to sift through perceptual data with the help of semantic knowledge and activations primed by past associations. There has to be some "place" where things come together for a given deliberative context, in which *the potentially relevant stuff* can be thrown against the proverbial cognitive wall to see what sticks. I have examined the proposals of Baars and Jackendoff, both of which present plausible models of how this global workspace might be characterized.

However, a Fodorian objection can certainly be raised at this point, that this is starting to sound exactly like the sort of non-local, unencapsulated processing that rules out modularity. The main problem a Fodorian will have with a global workspace is that it implies a level of global interconnectivity between modules that is implausible and unframed: how could it be the case that *every* module can post its output on the blackboard, and that every module can survey what's on that blackboard with an eye to inputs it can take up for processing? Indeed, it can't be quite that simple. Carruthers (2006a) argues that all systems do in fact "globally broadcast" their results, though this does seem to invite an intractable chaos of postings in the workspace. I think the key to answering this sort of objection lies in returning to the issue of memory retrieval and the limits of recall, as discussed in the previous chapter. One way to construe the global workspace is to highlight Jackendoff's characterization of it as a "workbench of memory" and marry that image to

Tulving's account of memory as reconstruction. The global workspace would then be operational only in cognitive acts that involve recall—either of semantic knowledge, or previous events, including belief and reasons for prior belief formation. Once recall is engaged, all and only those systems (or modules) that are activated by the retrieval context will be "connected into" the workspace at a given time. Given that the connections at the earliest stage will be brute associative ones, there is a likelihood that many will turn out to be relevant.

Perhaps a useful image would be to think of the workspace as a sort of switchboard, but with a very simple "operator" that essentially calls an *area code* and then lets whatever picks up the line try to communicate with one another—some will and some won't. The ones that won't or can't will hang up, or been hung up on. The ones that can (fruitfully) talk to one another will continue to do so. Presumably if certain lines of mutual communicability are consistently activated, a hard connection, or "hotline" could be forged between them. Certain intercommunications and co-activations are plasuibly so common that the "hotline" can form in such a way that there is no need to even connect those modules occurrently via the workspace: they can form an (unconscious, unmediated) interface. From this, a modular assembly can be seen to be emergent from constant co-activations of component modules. The key here is that the process *begins* will a call *out* from the global workspace—a recall cue—and that call out has a contextualized address, which limits the number of systems answering the call. Of those that do (or can), subsequent interaction will inhibit or disinhibit further activation.

This process can translate very well to belief revision: if the conscious task at hand is "check belief β" then the recall cue will automatically prime associated items in memory (and inhibit unassociated items). As I have already argued in the previous chapter, this actually rules out true Quineanism, as the nature of recall likely *won't* allow us to compare belief β to *all* (in principle) relevant belief, but only to the much smaller subset of beliefs constituted by concepts that are already organized or filed in associative relations. However, the process will certainly bring *many* items relevant to β into the workspace, and will activate many relevant modular processors to help sort and sift them.

Another helpful image here might be to recall the scene from *Apollo 13*, when the CO_2 levels in the space capsule are rising, and the crew does not have the right sort of filter

on board. At mission control, a call is put out to engineers in all departments who might have some idea about how to fix the problem, and a group is quickly assembled in a room to jury-rig a contraption to get the job done. The solution (a kludged-together "adaptor" designed to "fit a square peg in a round hole") is certainly inelegant, but it works. The reason I say the image could be helpful is that it mimics the sort of process, under description in this section, of a global cognitive workspace. It *isn't* the case that all sub-departments at NASA are constantly reporting in their work to one massive conference room, and when a problem arises, someone or something sifts through all that to find what's needed. And it *isn't* the case that some global central NASA executive knows what all the departments know and can put together the necessary information to solve the problem. It *isn't* even the case that some central NASA executive knows exactly *whom to ask* to sort out the problem. In actuality, all that is put out is a call that *names the problem*—in this case CO_2 filtering. At which point, all self-identifying relevant subsystems answer the call. From there, depending on the interactions of those subsystems once brought together, certain sorts of expertise find resonance and quickly build on one another in what can only be described as a sort of inter-departmental abductive reasoning process that aims at satisficing (rather than perfection). And note, this doesn't imply that all sub-departments are constantly at the ready, just waiting and listening for a call that they can answer: rather, certain calls can simply be salient in a way that grabs attention—just as when one's name, when called, instantly grabs one's attention, even when one was not actively listening. An identical process could be what goes on in belief revision: by consciously naming the problem, associative priming will alert various subsystems that they may be needed, and assemblies can fire up if their subcomponent modules are co-activated.

Granted, these are just images I am introducing to help illuminate how a global workspace could operate in ways in that don't invite various Fodorian concerns, and don't sneak the frame problem in through the back door, while attempting to screen it from the front. In the next section, I want to look at a concrete application of this sort of blackboard architecture: the IBM machine 'Watson', which is programmed to play the game show *Jeopardy,* and recently defeated the all-time (human) Jeopardy champion Ken Jennings. Watson is constructed using many of the design principles under discussion in this dissertation, and performs in many ways that seem to approximate human deliberation and

abductive inference. The success of Watson support the strength of the arguments I am defending that a massively parallel suite of modular processors can tractably approximate global—even Quinean-seeming—reasoning tasks.

6.1.4 Watson

Watson approximates abductive reasoning by running a massively parallel data-sifting system to find the 'best' answer to questions that often require cross-domain mapping (such as questions that involve puns or odd connectives). Some of the answers Watson is capable of are quite impressive. One example is "A *Green Acres* star goes existential (& French) as the author of *The Fall*." Watson nailed it perfectly: "Who is Eddie Albert Camus?" (Thompson 2010: 6).

> Watson uses more than a hundred algorithms at the same time to analyze a question in different ways, generating hundreds of possible solutions. Another set of algorithms ranks these answers according to plausibility; for example, if dozens of algorithms working in different directions all arrive at the same answer, it's more likely to be the right one. In essence, Watson thinks in probabilities. It produces not one single "right" answer, but an enormous number of possibilities, then ranks them by assessing how likely each one is to answer the question (Thompson 2010: 4).

Let's look a little closer at how Watson pulls that off, according to the design team, Ferrucci *et al.* (2010):

> The system we have built and are continuing to develop, called DeepQA, is a massively parallel probabilistic evidence-based architecture. For the *Jeopardy* Challenge, we use more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique we use is how we combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, or speed. (Ferrucci *et al.,* 2010: 68)

There are essentially 4 main design principles at work in Watson, paraphrased below (*ibid*):

Massive parallelism*:* Watson considers multiple interpretations and hypotheses, utilizing a multitude of parallel systems, each programmed to pick up on individual lexical cues in the question context.

Many experts: Various "expert" subsystems propose their analyses, and these are integrated in layered processing stages. (Note: a fine example of "blackboard architecture".)

Pervasive confidence estimation: Separately arrived at content interpretations are compared, and individual confidence levels are taken into consideration. Ferrucci *et al*. say that subsequently "an underlying confidence-processing substrate learns how to stack and combine the scores" (68). It's not clear how that stage works – perhaps it is simply a trade secret, but they don't elaborate much. Presumably it's some Bayesian-esque system that revises confidence intervals based on previous successful and/or unsuccessful responses.

Integration of shallow and deep knowledge: Lexical items are processed according to both "strict semantic" interpretations, and looser ones, and then compared in light of other results from parallel processing. This is akin to the semantic priming that Jackendoff discusses, and similar to what was mentioned regarding Bock & Lobell's research: all semantic connections get considered (literal, polysemous, ambiguous, known metaphorical uses). The interpretations that rise to the top following confidence estimations are the ones that will help formulate and filter the "final" answer.

To get to the "final" answer, Watson needs to "soft filter" the possible responses down to a tractable number for further processing (i.e., more evidence checking and confidence re-estimating) from that subset. For Watson, the soft filtering stage is designed to allow 100 candidate answers through the soft filter, but this is a "parameterizable function", according to Ferrucci *et al.* (2010: 71). Note that 100 is a lot more than a human can synchronously manage for compare/contrast purposes—*human* deliberation filters (somehow) the candidate answers down to something more like single digits before proceeding with deliberate analysis—so Watson has the advantage in that regard. Nevertheless, once Watson has whittled the choice set down to the 100 "best" answers, Watson *re*considers them, and searches for additional supporting evidence to test the 100 hypotheses.

> To better evaluate each candidate answer that passes the soft filter, the system gathers additional supporting evidence. The architecture supports the integration of a variety of evidence-gathering techniques. *One particularly effective technique is passage search where the candidate answer is added as a required term to the primary search query derived from the question. This will retrieve passages that contain the candidate answer used in the context of the original question terms*. Supporting evidence may also come from other sources like triple stores. The retrieved supporting evidence is routed to the deep evidence scoring components, which evaluate the candidate answer in the context of the supporting evidence. (Ferrucci *et al*. 2010: 72, emphasis added)

Notice the feedback mechanism in place, "where the candidate answer is added as a required term to the primary search query derived from the question ... [to] retrieve passages that contain the candidate answer used in the context of the original question terms." This recursive feature is one of the key elements in the process, and one that human evidence-checking likely employs also—it's a way to bootstrap up confidence levels using Bayesian procedures. One aspect of Bayesian updating is that "today's priors are yesterday's posteriors" (Lipton, 2004: 115)[6]—but "yesterday" is just a metaphor here: one split-second ago's posteriors are this split-second's priors also. A Bayesian inference machine faced with an elaborate data set and a quick search algorithm can easily bootstrap up a confidence level by re-examining the evidence in recursive cycles with each iteration increasing the likelihood, and hence giving more weight to confirmatory evidence in the set, assuming there is no time to check *all of it.*[7] Once we have a hypothesis in place that has already been soft-filtered and comes with a presumption of relatively high confidence, the evidence *is re-evaluated with this hypothesis in mind*. This can help limit or speed up evidence searching in a couple of different ways. First, it can laser focus the search on tighter bands of evidence, as the search narrows to confirm certain assumptions. And secondly, it may allow for some bit of isolated evidence, which had been previously swamped and hence unnoticed, to come newly into view as it survives the whittling down function, or that, alternatively, is recalled under the revised recall cue whereas the initial recall cue, being more broad, may have somehow inhibited it. For Watson, the re-evaluation stage is not explicitly *confirmatory* in nature—though the evidence search will be limited to items that resonate with the "new" revised question, so one could argue that a *de facto* confirmation bias sneaks in.[8] For human reasoning, it seems more likely that an *explicit* conformation bias is "programmed" into the process, to expedite matters (even at the cost of a higher rate of incorrectness—given that a "correct" answer arrived at too late is essentially a "wrong" answer, in *Jeopardy* and in life).

So let's take stock here and notice the number of ways that Watson's design principles coincide with the *global workspace* and *parallel architecture* proposals we have looked at in this chapter.[9] There are a number of interesting aspects to Watson's design that seem relevant to the discussion of human context framing and inferential practice. First and foremost, Watson operates via natural language. This is not to say his *programming* operates via natural language (presumably, he has a machine code syntax for processing—his own

inner mentalese). But he interacts with information via natural language, and he is designed to run his informational searches through the medium of natural language. As Ferrucci *et al.* note, this means that there will be difficulties right from the start, as "questions and content are ambiguous and noisy and none of the individual algorithms are perfect" (2010: 67)—hence Watson's design principles diverge from standard Boolean processing, and instead must resort to heuristic, massively parallel processing systems with interface mechanisms. Forcing the machine to operate in a problem domain of human language entails getting the machine to think more like a human might. This parallel architecture converges in something just like a "blackboard architecture" where a collection of encapsulated domain-specific "experts" have thrown their "results" into the mix, and statistical inference generators assess probabilities based on overlapping conclusions—what Jackendoff describes as "resonances" (or perhaps what even Fodor describes as conceptual "locking", for all the metaphysical weirdness of his "whirlpool" attractor landscape). All this to say, Watson's design is remarkably similar in many ways to the *human* "designs" under discussion in this and the previous two chapters. And Watson's success at a very "human" game like Jeopardy suggests *support* for those cognitive design principles as being the same as, or at least very similar to, our own.

Of course, despite all this effort to understand our own intelligence based on a certain machine intelligence design is sidestepping a very common belief that the analogy is flawed from the start. A common, standard dismissal of artificial intelligence is that it isn't "intelligence" at all, but mere mimicry.[10] Here's Stanley Fish responding to Watson's *Jeopardy* win:

> It's just a bigger and fancier version of my laptop's totally annoying program. It decomposes the question put to it into discrete bits of data and then searches its vast data base for statistically frequent combinations of the bits it is working with. The achievement is impressive but it is a wholly formal achievement that involves no knowledge (the computer doesn't know anything in the relevant sense of "know"); and it does not come within a million miles of replicating the achievements of everyday human thought. Watson's builders know this; when they are interviewed they are careful to stay away from claims that their creation simulates human mental processes (although they also murmur something about future hopes). But those in charge of the artificial intelligence hype are not so careful and they delight in exciting us and frightening us with the fiction of a machine that can think. It's great theater, or in Watson's case, great television, but that's all it is. (Fish, 2011)

I think one of the main reasons arguments like this are made is that they are working from a presupposition or definition of *human* intelligence that is flawed—an idealized, Quinean, interpretation of human inferential capacity. I think it's actually quite correct to say that what Watson (or any AI) is doing is "simulating" or mimicking or otherwise *approximating* human deliberative processes—but I would qualify that by noting that *I think this is also true of humans.* I think *we too* merely approximate or "simulate" what we (ideally) construe reasoning to consist of. As I have argued throughout this dissertation, the "relevant sense of 'know'" that Fish mentions above is precisely a sense of 'know' that we almost certainly don't live up to: the *finitary predicament* rules that out.

Recall that in formulating the frame problem and its attendant pessimism about ever understanding the central systems of human intelligence (e.g., Fodor's "First Law"), one of the pieces of evidence that is held up in support of that pessimism is the failure to design intelligent, modular AI programs—the argument being: if *we* are just intelligently programmed machines—nested hierarchies of modular processors—then why haven't we replicated that success mechanically yet? However, if more and more successes such as Watson are built, modeled on assemblies of modules running parallel heuristic algorithms and integrated via semantic associations, statistical hypothesis evaluation, inference generators, and parametrizable confidence thresholds, then arguments supporting Fodorian pessimism collapse. If we can create seemingly deliberative processes mechanically, then we can explain how it is that our own intelligence need not court frame problems, or ghostly homuncular processes: every process remains a local one, and the process is *dumb,* not deliberative. In the following section, we will examine some promising suggestions as to how *we* do it—we will look at the evidence regarding the heuristics and biases that human minds rely on in order to reason tractably. After which, I can repose the question: will these heuristics and biases, combined with a global workspace/blackboard architecture, be *enough* to approximate the sort of global, holistic, seemingly isotropic processes that rational inference and belief revision practices at least *should* have in the ideal (Quinean) sense?

6.2 Heuristic approximation

In §1.3.4, we already looked briefly at 'dual process' theories, all of which posit the existence of various reasoning heuristics and cognitive biases that allow for speedy efficient

processing specifically via *bypassing* reflective, conscious deliberation. Recall Stanovich & West's point that these "System 1" reasoning processes "automatically contextualize problems"—what is referred to as the *fundamental computational bias*. An automatic contextualization of problems, of course, is precisely what is needed to avoid the frame problems we have been discussing, including the sorts of problems one might pose regarding the identification and coordination of information relevant to whatever task is currently featured in the global workspace. Heuristics and biases—pre-programmed, unconscious, mandatory operations aimed at simple satisficing, rather than maximal or optimal solution generation—are the best possible answer. Recall Gigerenzer & Todd's (1999) point that what we need are heuristic algorithms that can aid in search and judgment procedures (i.e., where to start, when to stop). With appropriate heuristics in place, we should be able to *approximate* rational practice—something *good enough*, though not perfect; tractable, yet prone to systematic patterns of breakdown, cognitive illusions, and incorrigible, unconscious biases as a result.

6.2.1 Heuristics & biases

The classic studies in the "heuristics and biases" research program are from Tversky & Kahneman in the 1970s—the original 3 heuristics identified in those studies are 1) *representativeness*, 2) *availability*, and 3) *anchoring and adjustment*. Much of the heuristics and biases research takes on board the assumption of dual-process theory—System 1 and System 2—where the former operates reflexively, automatically, computationally frugally, associatively, un- or sub-consciously, quickly, and skillfully, while the latter is reflective, controlled, computationally demanding, inferential, conscious, slow, deliberative, and rule-governed.

We already looked at the example, in chapter 1, of what Tversky & Kahneman call the *representativeness* heuristic and how it works to color our perceptions by way of stereotypes. Recall how in the *Linda the feminist bank teller* study (Tversky & Kahneman, 1983), subjects routinely are led to commit the conjunction fallacy in their reasoning about whether it's more likely that Linda is a bank teller or a feminist bank teller. In that case, the explanation of the reasoning "mistake" is that we (unconsciously) latch onto a stereotype of certain sort of "progressive" person, based on certain salient cues in the information we are

asked to read—once the stereotype is called up, we reason according to *it* rather than to the specific details we are actually reading, and apparently without consulting our "logical" faculties to double-check the response. A stereotype "bell" is rung, and a judgment is called based on it, even if in this case it is an *illogical* bell. The decision (that 'feminist bank teller' should be ranked more likely than 'bank teller') is *quick*, and indeed, the deliberative processes necessary to understand why it's incorrect actually take conscious, time-consuming deliberation in some cases.[11]

Representativeness also plays a role in subject's understanding of randomness and base rates—statistical understanding in general—as the representative stereotype of what we think, for example, "random" means may cause poorly informed judgments of *non*-randomness in perfectly random situations (e.g., a coin flipped 5 times may turn up heads each time, *and still be random,* but we won't be able to perceive that as randomness, since it doesn't fit the stereotype). The "sympathetic magical thinking" discussed in chapter 1 with reference to studies by Rozin *et al.* would be another good example of the representativeness heuristic driving aversive responses to things we need not actually be aversive towards (like feces-shaped chocolate). This is what heuristics like representativeness can do: drive us towards quick unreflective judgments. Most of the time, they do us a favour in this regard—conserving time and cognitive resources—though they reveal themselves in reasoning "mistakes" when they activate automatically in contexts where it would be more profitable to *avoid* employing them.

The *availability* heuristic is one that prioritizes the most "psychologically available" events in memory associated with a particular deliberative context. Of course, these will often *not* be the most relevant events. A classic example is the fact that many people express some trepidation about flying in airplanes, despite knowing that, statistically, air travel is safer than car travel. There are many similar examples (fear of terrorism vs. tornado, fear of shark vs. being hit by a falling coconut, etc.) in which highly *publicized* events come more easily to mind when assessing risks, although the mere fact that they are highly publicized is probably a good indicator that the events in question are relatively *not* as frequent. Our judgments of risk are often highly sensitive to personal experience or exposure—reasoning via anecdote is extremely common. And extremely available information will find its way into our judgments even when it's patently incorrect. My favorite example to give to

students is the following question: "Timmy's mother has three children, their names are Snap, Crackle, and _____?"[12] Tversky & Kahneman (1973) explain that the availability heuristic can lead to self-reinforcing feedback loops: events that provoked strong responses (e.g., fear, anxiety) will by virtue of that fact be more readily available in the *next* fear or anxiety-provoking context, and then the fact that they come easily to mind will in turn reinforce their importance and availability in future judgment contexts. We will see in PART III, below, this will be a serious issue in memory retrieval insofar as "remembered" misinformation can be self-reinforcing, leading to rich false memories, and potentially unrevisable beliefs, including delusional belief.

The *anchoring and adjustment* heuristic (Kahneman & Tversky, 1974) is used in estimating quantities or values: a known quantity (or value) is used as an "anchor" and then adjustments are made away from that anchor until a satisfactory value is reached. Often, this can be a highly useful estimation tool—for example, as I write this paper, I am wondering if I have time to finish this section before dinner. I recognize that I have been writing this section for about an hour, and I am on the third of the three heuristics I will discuss: so I anchor to *one heuristic/30 min* and then shave off 10 minutes, since I've already started writing about this (final) one. I conclude I'll be done in 20 minutes—in time for dinner. In *this* case my estimate will probably be pretty close, since my anchor is relevant and clear. However, in many instances, the anchor we start with may be poorly chosen (not representative or relevant; based on too little information; mistaken; too far away from the value we are looking for) and the resulting adjustment will fail to come close to the correct answer. For example, if I were asked to answer the following: "A person who earns $35,000 a year is in the top __% of worldwide income per capita?" I would make my estimate on two things: my own income, and my estimation of what percentile I think I am in, relative to the world. In this case, I would probably end up guessing "top 10%", and I would be off of the correct answer by a factor of 10.[13] In this case, my anchor is too far off the reality (I am in the top 1%, and yet didn't know that—since it generally doesn't *feel* true, especially given North American political arguments which reference "the top 1%" as a very small group of very rich people, relative only to North American standards). As a result, my adjustment fails completely, because the starting place is flawed.

The other way in which this heuristic can lead us astray is when the anchor is actually a good (relevant) start, but the adjustment away from it is insufficient—either because of a faulty understanding of the breadth of whatever continuum is being employed, or because the starting point was simply too far away from the answer. Epley & Gilovich (2006) explain that "adjustments from self-generated anchor values tend to be insufficient because they terminate once a *plausible* value is reached" (311, emphasis mine). In the income example, my answer of 10% was plausible enough (indeed, technically true), but is not really very close to the appropriate or relevant answer.

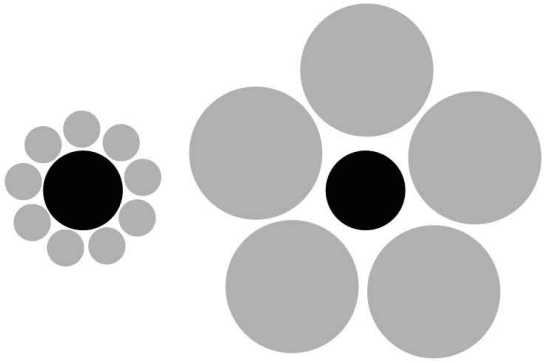
Anchoring and adjustment also leads to *numerous* other identifiable reasoning biases. Indeed, Gilbert (2002) argues that the anchoring and adjustment heuristic, is the "obscure sibling" (insofar as fewer studies have been published on it, compared to the other two "celebrity heuristics"), despite the fact that anchoring and adjustment "may well be the one that psychologists not yet born will consider the most important" (Gilbert, 2002: 167).

> [I]t describes the process by which the human mind does virtually all of its inferential work... judgments are generally the products of non-conscious systems that operate quickly, on the basis of scant evidence, and in a routine manner, and then pass their hurried approximations to consciousness, which slowly and deliberately adjusts them" (Gilbert, 2002: 167).

Ariely's (2008) *Predictably Irrational* explains how the anchoring and adjustment heuristic doesn't just drive our inferences about *numerical* or quantitative values, but about *values in general*—the prevalence of this heuristic in our reasoning serves to "relativize" almost all of our deliberations to key anchors. He argues that in general, "most people don't know what they want until they see it in context" (3). However, the anchors we use to "locate" our preferences are often arbitrary and/or irrelevant in the grand scheme of things, leading to incongruous and intransitive preference formations. A standard example is the following, adapted from Tversky and Kahneman (1973): most people would be willing to walk an extra 15 minutes to buy a $25 item if they were told it was on sale at the farther location for $18. However, when about to plop down $450 on an item at one store, the news that it's on sale just a 15 minute walk away for $442 is *not* likely to cause most people to make the trip Why? Because when anchored to $25, a $7 savings feels pretty good! Whereas if your local anchor is $450, suddenly $8 is not worth your time. That makes no sense in terms of rational preference, of course. $8 is objectively worth more than $7—but we don't value things

objectively. Nor do we respect transitivity: \$7>15 minutes; \$8>\$7; 15 minutes>\$8? That's not rational.

Ariely compares the act of valuing and judging as prone to error in precisely the way perceptual systems are prone to illusions. He gives the following visual illusion, and explains the analogue in the quote that follows:



> As you can see, the middle circle can't seem to stay the same size. When placed among the larger ones, it gets smaller. When placed among the smaller circles, it grows bigger. The middle circle is the same size in both positions, of course, but it appears to change depending on what we place next to it. This might be a mere curiosity, but for the fact that it mirrors the way the mind is wired: we are always looking at the things around us in relation to others. We can't help it. This holds true not only for physical things—toasters, bicycles, puppies, restaurant entrees, and spouses—but for experiences such as vacations and educational options, and for ephemeral things as well: emotions, attitudes, and points of view. (Ariely, 2008: 7)

Ariely has done numerous experiments using anchoring and adjustment to set up what he calls "decoy effects"—where the presence of an *unwanted* option in a choice set can affect the preference judgment (2008: 5-6). For example, given the following 3 options for subscriptions to the *Economist* magazine:

A. Online access only - \$59/year
B. Print only - \$125/year
C. Print and online access - \$125/year

The preferences of the student groups he administered the choice set to were clear: no one wanted the dominated option (print only for \$125), while the students split between the two remaining options, with 84% opting for the combined deal (print and web for \$125). When other groups of students were offered the choice with option B *removed*, however, the preferences changed dramatically, with option A suddenly becoming the majority preference (68% chose online access only). Ariely's conclusion is that option B acts as a "decoy", in

that it makes option C suddenly more preferable.[14] The reality is that the students clearly aren't sure how much they want the *Economist*, and whether they prefer print or web access, or both. But the fact that print alone is anchored at $125, and online alone is $59, suggests that print+online should be something like $184, so that $125 combo deal sounds like a "good deal", and draws preferences accordingly. Marketers of all stripes use this kind of anchoring to direct our valuations: the $60 bottle of wine on display makes the $40 one look reasonable; the charitable donation form has pre-filled amounts to tick off asking whether I'd like to give $50, $100, $200; the car salesman gives me 3 months free subscription to the satellite radio network, even though I walked in having previously decided that was an unnecessary option.[15] Ariely even ran a study to determine that physical attraction is relativized to anchors (2008: 12-13). He distributed photographs of 3 people, two who looked relatively similar, though one was noticeable more attractive, and the third was completely different looking. The majority of subjects rated the better-looking of the two similar faces as more attractive, as they had something to anchor to—one of the pictures couldn't be compared as easily, *and therefore wasn't.* While of the two that could be easily compared, the "better" got the nod most of the time.

The important lesson in these sorts of studies is that our judgments are highly sensitive to contextual effects that may or may not be relevant. We have apparently adapted certain "rules of thumb" as Kahneman & Tversky call them, to come to judgments *quickly* and without much, or even any, conscious "thinking". In the complicated modern world, these may "misfire" quite often, and lead us to sub-optimal judgments—or just plain incorrect or irrational ones. However, most of the time, we reach the point of "satisficing," to use Simon's terminology: note that even when we *do* make unreasonable judgments via heuristics, we *still* tend to "feel" pretty good about them, and often dispute the idea that we are mistaken. For example, try convincing the students in the *Economist* case that a good third of them would have chosen differently with the mere inclusion of a dominated option that no one wants—they would likely find that preposterous. And yet, a good third of them would have done just that.[16] Similarly, the sheer illogic and preferential intransivity of opting to walk 15 minutes to save $7 one day, and then to subsequently determine a 15 minute walk for $8 to be idiotic the next day, hardly bothers most people who have it pointed out to them. It makes no sense, literally, but it *feels* like it makes sense when it happens.

Perhaps one reason for this is that a great deal of "sense-making", as an activity, is taken care of in System 1, mediated by heuristics and biases that operate below consciousness, more at a level of reflex than reflection. Even though, ideally, "sense-making" would be a singularly System 2, Quinean affair, the reality of quotidian sense-making is that System 2 is too slow, and System 1, despite its flaws, is fairly effective—and at least it gets things done.

But why evolve a system that makes mistakes?[17] Why would nature supply the human mind with judgment heuristics that are so sub-optimal? The answer to that is, of course, the *finitary predicament*—we have no choice but to settle for heuristic approximations: we don't have the time or resources to actually work things out optimally. We already discussed in chapter 2, above, Gigerenzer & Todd's account of "fast and frugal heuristics" bequeathed to us genetically as part of an "adaptive toolkit". I will not re-run that exposition in depth here. Recall, simply, the claim that empirical evidence suggests the near-universal possession of deliberative heuristics divided into 3 main categories: those designed for 1) guiding searches; 2) stopping searches; 3) decision-making (1996: 16-17).

> Fast and frugal heuristics employ a minimum of time, knowledge, and computation to make adaptive choices in real environments. They can be used to solve problems of sequential search through objects or options, as in satisficing. They can also be used to make choices between simultaneously available objects, where the search for information (in the form of cues, features, consequences, etc.) about the possible options must be limited, rather than the search for the options themselves. Fast and frugal heuristics limit their search of objects or information using easily computable stopping rules, and they make their choices with easily computable decision rules. (G&T, 1999: 14)

6.2.2 Test case, or: how I learned to stopped worrying and love my new TV

Let's take a moment to run through a fairly typical example of deliberation, in order to highlight how System 1 and 2 processes interact to frame the deliberative task. Imagine I need a new television set: which one should I buy? In the description of how heuristics and biases rule System I 'reasoning', there is an interesting dance going on between conscious and unconscious processes. On the one hand, when I am deciding which new TV set I "need"—the 27" Sony, the 50" Samsung, or the 55" LG—I am fully *conscious* of my deliberations. I don't just walk in the store and walk out with a TV automatically, with no recollection of how that happened. When I shop for a TV, I *feel* like I am actively (carefully, consciously) deliberating. However, as we have seen, my deliberations are very likely

framed heuristically—the reality is that I probably have no idea what size or brand TV I "need" (or whether I need one at all). TV's are not directly associated with any "need" that my brain is adapted to.[18] But I *want* a TV, and I need to fulfill *wants* (to some satisfactory degree), and my brain helpfully frames the task for me with fairly simple heuristics: the availability heuristic will prompt salient recent memories (e.g., my neighbour got a Samsung and said the picture wasn't bright enough), the representativeness heuristic will impose certain assumptions accordingly (e.g., Samsung TVs aren't bright enough), and the anchoring heuristic will help quantify my "need" (e.g., the majority of the TVs on display are large, so I'll start looking at that class, and select the best). I walk out of the store with a new 55" LG TV, and a sense of accomplishment—I made a good, thoughtful, deliberate decision.

Of course, that's partially a lie—given the unconscious action of the three heuristic framing mechanisms, I unconsciously limited my deliberation in ways that on reflection seem downright foolish. I *should* know better than to think that my one anecdotal piece of evidence about my neighbor's Samsung is meaningful. I *should* recognize that his set may in fact be entirely *un*representative of Samsung TVs in general. I *should* realize that I never looked at any statistically significant sample of Samsung owner satisfaction levels, or any bench tests comparing the brightness of the Samsung to other brands. Perhaps my neighbour is right about the Samsung not being bright—nevertheless, I *should* realize that it *may still be brighter than the Sony and the LG!* And I *should* notice the arbitrary nature of my anchoring to the size. If the store had alternatively displayed three additional 27" TVs, would I have still looked exclusively at the larger ones until I found a "good enough"? Or would I have focused on the shelf with the smaller sets, given that they were more common, and found a "good enough" one there, without ever considering the 50"+ TVs?

My decision was "conscious" and "deliberate", but circumscribed by unconscious, and undeliberative processes that seriously and somewhat arbitrarily limited the extent to which my conscious deliberation was employed. This supports the idea that heuristic, automatic processes pre-limit and frame deliberative tasks to impose frugality and make problems tractable. However, what about the fact that I can *see* all this in retrospect, and that I could, if I had *taken the time*, moved my TV deliberations into System 2—noticed and consciously avoided the heuristic approximations and focused instead on explicit, coherent logical reasoning? I certainly am capable of it—I just did it in the previous paragraph. But

what are the systems that *allow* for that reasoning to happen, in a tractable way? As an answer, I will return to the claim stated at the end of the previous section, I think that no matter what, even the most deliberative, conscious, actively assumption-avoiding, prototypical System 2 reasoning processes end up constrained by some lower-level System 1 work—specifically in one area: memory retrieval. As we have seen, recall from memory is largely associative, compartmentalized and circumscribed by the *encoding specificity principle*, and prone to sub-optimal breakdowns, including possible revision of memories during retreival, and the inhibition and even unrecallability of others, depending on the level of congruence between encoding and retrieval contexts. Even when we *try* to be exemplary Quineans, our practice will be limited by the underlying modular processing of belief organization and recall. Careful, deliberate System 2 reasoning can certainly approximate the Quinean ideal, as we can perhaps force attention to search farther than might be the case in System 1 processing—but that slower process is largely the result of simply running more, and perhaps repeated, System 1 processes. And all of that can be instantiated by a suite of massively parallel processing modules.

6.2.3 Is there a frame problem regarding heuristic selection?

Of course, even with this account—where cognitive resources are conserved by running "good enough" subroutines, honed by selective pressures—we find, once again, a lurking version of the frame problem. In this case, the problem rears its head if we ask how the cognitive system (on the whole) "knows" or recognizes or delineates the deliberative context at hand in such as way that the *appropriate heuristic* is employed. If the story is that, when faced with a time-consuming deliberative task, we engage our System 1 resources, and automatically go to a rough and ready rule of thumb, then we have just invoked a second order deliberation—namely, which rule is the appropriate one—and perhaps courted a framing regress.

Fodor strenuously objects to the idea that heuristics could approximate global deliberation for precisely the reason mentioned above: 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*. Fodor calls out Carruthers for inviting this regress, specifically, as Carruthers invokes the possibility that ''choices [among heuristics] could be made by higher-order heuristics, such as 'use the one that worked last time'" (Carruthers, 2001: 30). Fodor complains that if this is the case, we still need a system that determines which previous contexts the present one is similar to, so that we can judge which heuristic 'worked last time'.

> What I'm to do when I'm *doing again the same thing that I did before* depends on what I'm to take as a recurrence of *the same situation I was in before*, either in general or for the case in hand. But that, in turn, depends on what I'm to take to be the *relevant* description of my previous situation; the description under which the kind of action I performed explains the success of my action. So, lacking an account of relevant sameness, the advice 'it worked last time, so just do the same again' is empty. (Fodor 2008: 119)

I would dispute Fodor's claim here that we are "lacking an account of the relevant sameness" of situations. Given that previous situations are *recalled*, and given everything we have discussed about the encoding specificity of memory, and how recall cues associatively prime contextually congruent memories, then the ability to determine "relevant sameness" between situations hardly seems like an issue: it doesn't really take any "thinking" at all. Relevant sameness can be specified by the organization and compartmentalization of memory in an automated and subdoxastic fashion. Sometimes we might *miss* an instance of relevant sameness, or work from an instance of sameness that isn't the most optimal one, or inadvertently end up working from a presumption of relevant sameness that isn't actually appropriate at all. The point is that it will *mostly* work. And that's all heuristic processes aim at. So a simple, *nested* set of heuristics, including a couple of general heuristics that direct the activation of more specified heuristics, seems highly plausible.

Another very plausible response to the Fodorian objection is that, given heuristics ostensibly evolved in response to selective pressures, those same pressures, or associated ones, will *automatically* engage the corresponding heuristic, reflexively—without even invoking the higher-order heuristic selection heuristics Carruthers employs. Goodie *et al.* (1999) argue long these lines, noting that heuristics that *didn't* automatically kick in as

needed, without thinking, *in the relevant contexts,* could never have evolved in the first place. It does seem an odd concern, on Fodor's part: if his objection holds, it should apply to *all* domain-specific mental processes, right down to the most basic level. If there is an "input problem" regarding heuristic "selection" then there presumably should be one when it comes to *reflex* "selection", as well. Couldn't we ask the same sort of question about *ducking?* To borrow Fodor's question, "isn't deciding how to solve a problem itself a species of problem solving?" Well, in the case of ducking, how is it "decided" that ducking is the appropriate reflex to engage, rather than bracing, or blinking, or running, or laughing? Of course, the answer to this is that certain environmental features simply trigger the ducking reflex automatically—there is no need to invoke a decision stage between the situation and the triggering of the reflex. But then why not accept the same solution with regards to heuristics that aid in deliberative tasks? I suppose the objection will be that such contexts are not quite as *brutely* "obvious" or uniform as the contexts that require ducking. However, it may be pretty simple: for example, anchoring and adjustment could be automatically engaged in all quantificational/evaluative contexts. Period. That's certainly what Gilbert thinks, as noted above. Similarly, we could turn again to one of the suggestions looked at earlier in this chapter, such as Barrett's "enzymatic" analogy—perhaps heuristics actively "seek out" problems that fit them, or at least are primed to "take the call" when a relevant problem domain is activated. This seems of a piece with the ideas mentioned above by Goodie *et al.* and Gilbert: we don't need to "choose" heuristics appropriate to a task—rather, heuristics blindly "choose" the tasks they engage, via proprietary activation, just like a sensory module or reflex, or as Barrett suggests, an enzyme.

Furthermore, there is no reason that the selection of heuristics cannot be, to a certain extent, *arbitrary*. As long as the *satisficing* policy is assumed, 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. Samuels echoes the point that I have made repeatedly regarding the (seeming) globality of reasoning which is relevant to this distinction, arguing 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, in a normative sense, doesn't mean that we *actually* satisfy the exhaustive demands of that globality in the real world.[19] We don't need to be Quineans about heuristic selection, of all things. And even if we wanted to be, the list of heuristics is a fairly *short* list, which doesn't require a lot of background searching for relevance at all.

I think, therefore, that a Fodorian sort of "input" objection aimed at heuristic selection in decision tasks is misplaced. As we have seen, there is voluminous evidence that heuristics are used in reasoning and decision tasks, usually outside of conscious awareness, mandatorily, and prone to systematic breakdowns. The sort of "predictable irrationality" that Ariely speaks of, compares common reasoning "mistakes" resulting from the (mis)use of heuristics and biases to the "mistakes" of perception we find in illusions. Indeed, the argument would be that optical illusions involve unnatural 2D manipulations of the perceptual scene in a way that our modular, encapsulated senses represent incorrectly by imposing the *usual* assumptions in an attempt to reverse-engineer the illusion to a representation of a 3D world. And in an exact analogue, *cognitive* "illusions" arise when the environment to which our heuristics and biases were adapted is not quite the same as the environment in which they now get deployed. The sort of reasoning biases and cognitive illusions uncovered and studied in the heuristics and biases research program is highly evocative of modular processing with regard to the characteristics mentioned above. Just as perceptual illusions are taken as a telltale sign of modular perceptual systems, so too should cognitive illusions and reasoning biases be taken as the telltale signs of modular reasoning systems.

This is not to say, however, that I am suggesting that certain heuristic "rules" reside in some sort of module—that the *representativeness* heuristic, for example, *is a module*—that wouldn't even make grammatical sense. What I would suggest, rather, is that the *representativeness* heuristic is a second-order description of the *sort* of processing that would be entirely mediated by modular processors. It's not that there is a heuristic algorithm, there in the brain, waiting to be called upon by various systems—rather, *various systems independently* run algorithms that are constrained by principles that fall under the banner of that particular heuristic. The "rules of thumb" are not rules that are *turned to*, they are rules that systems are apparently *designed by*. Heuristic processing is inherent in deliberative

systems—baked in to the processing subroutines that underwrite the system's global function—it's not *another* system that is called in as aid.

6.3 Dumb as a bag of hammers

For the above-mentioned reasons, I think that the Gigerenzerian image of heuristics as part of an "adaptive toolkit" is somewhat misleading, as it suggests the image of a global tool "user" (and all the local frame problems that might entail). If we want to stick with the tool analogy, I think it might be more accurate to describe heuristic algorithms as merely a set of *hammers*. Imagine your toolkit consists only of hammers, perhaps of a few different sorts (sledgehammer, claw hammer, rubber mallet, ball-peen) and they come pre-tagged and roughly categorized to use on certain jobs. So when faced with a job, one simply pulls the hammer tagged for that job, and then, as the saying goes: *to a person with a hammer, everything looks like a nail.* Even if you aren't sure you are looking at a nail, you can try hitting it with a hammer: it that works, then it's (close enough to) a nail. That's what heuristics and biases do: they transform the deliberative landscape into so many nails, simply by supplying you with the hammers. Hammers may not be the most delicate or accurate or optimal tool for most jobs—they are pretty brute instruments—but a set of hammers will *do* in a large number of cases, and they're pretty cheap and easy to operate, in terms of expending resources, so it's a decent cost-benefit tradeoff.[20] And, running with the hammer metaphor, if nature hands you a cognitive hammer, and suddenly all your "problems" start to look like nails, then those problems just became eminently *tractable.* The tool both clarifies and transforms the problem space so that everything is either in a form that the tool can handle or it's ignored. The hammer frames the world quite effectively. So maybe we are actually *smart* as a bag of hammers?

However, despite the clear benefits of heuristic problem solving, the question of tractability still arises in another way: regarding System 2, and the occasions in which we *are* capable of going around, overriding, or correcting the biased, heuristically approximated judgments (beliefs) arrived at via System 1. The objection would go as follows: isn't the fact that we *can* notice our biased thinking and work around it evidence of *non*-encapsulated processing (so my account above is simply incorrect)? Even if System 1 is essentially just a description of a suite of interacting heuristically-driven modular sub-structures, System 2 still

must be something else, something non-modular and more Quinean in nature, in order to *notice* the (often sub-optimal and biased) heuristic processing at work in System 1. In other words, if the "bag of hammers" metaphor is taken at face value, *then how are we ever able to recognize it as such?* There are occasions where we notice that we are (unconsciously) hammering away at "nails", and that the hammer *isn't* actually the best tool for the job, and we carefully search for better, more refined tools to tackle the problem. Even if such occasions are few and far between, they do seem to happen *some* of the time, and, so the objection goes, these instances at least demand a more Quinean, non-heuristic, and non-modular cognitive system.

I am arguing that the answer is *no*—and that the answer *must* be no, because that is what the 'finitary predicament' demands. Even the "slow" reflective deliberation of System 2 is entirely subserved by System 1, which is in turn comprised of modular systems and constrained by heuristic algorithms—just in more complicated feedback and sifting arrangements, so that "access" *is* essentially more global, yet that global processing is nevertheless always constrained by locally processed compartmentalization of tasks (echoing Cherniak), and regularization routines that impose assumptions that aid in tractability. Modular assemblies, at multiple levels, make all of this possible:

- Sensory modules at the periphery represent raw perceptual data.
- Perceptual integration modules cross-check, run cross-modal error-correction routines, interpolate, and otherwise smooth perceptual 'scenes'
- Massively parallel dedicated cognitive modules sniff out "subtle cues" from those perceptual scenes and perform further integration of perceptual contents with associated semantic knowledge. Possible examples may include:[21]
    - social cognition modules can detect other minds (*ToMM*) or free-riders (*CDM*); conspecifics; potential mates; diseased individuals; etc.
    - threat detection systems can isolate specific dangers (snakes, heights, toxins, etc.) and automatically activate aversive behaviour;
    - language modules can integrate phonological, syntactic, semantic and pragmatic sub-systems, and appropriately interface with language production and associated motor systems;
    - face and object recognition systems can pick up on specific stereotypes to order and tag perceptual inputs by category for processing (and storage for future processing);

    - logical argument analysis modules can detect argument types and direct to subroutines (*modus ponens/modus tollens* argument evaluators; testimony evaluation; Bayesian simulators; etc.)

  Note: nothing need "direct" inputs to these processing modules, as I have argued—the modules can be domain-specific, and on the metaphorical lookout for appropriate input representations. When activated, they process according to the encapsulated algorithm. From there, outputs may or may not be picked up for further processing, or by conscious awareness.

- A global workspace allows for online interfacing of a limited set of items from across the cognitive economy—entry into the workspace at any given time is managed by heuristic search and activation functions, as well as the limits on recall imposed by the *encoding specificity principle*. Tractability can be further imposed by interface modules setting activation thresholds, gating access, and soft-filtering inputs to the workspace.
- Conscious "simulations" based on whatever is currently in the workspace can help bootstrap up confidence levels and prioritize items for workspace entry. Simulations could set expectations that then redirect (or "requery") other systems to confirm or disconfirm. Similarly, simulations can trigger various inhibition/disinhibition functions based on contextual or associative "resonances". In short, the problem can be iteratively reposed with updated assumptions—the job can be resurveyed with a particular tool in mind.
- Heuristic judgment algorithms set thresholds for satisfaction that halt processes when plausible, or "good enough" responses are hit upon.

For the most part, the "sifting" functions take place naturally via modular sub-processors that do the bulk of their task-minimization work largely with respect to highly domain-specific memory retrieval algorithms. The fundamental tractability issue with holistic thought is the exhaustive *memory* searching it ostensibly entails. Solve that problem, and the reasoning itself isn't an issue: we *do* have a logic faculty of some sort after all. It's slower, and we might bypass it a lot for the sake of speed, but it's there when needed. The only restriction on it is that it can only engage propositions that are currently before it in short term memory—in the workspace. Recall is the mediating step, and it is heavily circumscribed: memory isn't isotropic, after all, and recall isn't Quinean, even when *we* are trying to be. Exhaustive memory searches can't happen, and regardless, are not necessary: concepts and beliefs are organized in a compartmentalized fashion, and retrieval serves to promote

contextually relevant items and inhibit non-relevant ones by brute associative priming—*not* by "deciding" what's relevant and what's not.

6.4 Review and look ahead

On the central question of the dissertation—how do we frame deliberative contexts in order to manage and revise belief in a tractable fashion—a number of proposals have been examined, namely, the existence of a global workspace of working memory, the use of heuristics and biases in deliberation, and the interplay between dual systems of reasoning, including how both could be subserved by massively parallel, domain-specific modular structures. The resulting over-arching system could be considered *virtually* encapsulated as a result: no individual part of the parallel system is unencapsulated, and the inter-system activations are all mutually domain-specific, so the emergent interaction system inherits tractability. I have noted my own revision to the "adaptive toolkit" argument regarding heuristics and biases, suggesting that a better metaphor would be simply a *bag of hammers*, given the cliché that "to a person with a hammer, everything looks like a nail"—the tool literally may transform the perceptual environment in a way that lets it do its work. Heuristics and biases transform our problems into formats that can be tractably managed. The result is brute, but effective—"good enough" but prone to be sub-optimal.

This sub-optimality will reveal itself in a number of ways. The simplest way it will be revealed is in pervasive *unawareness of inconsistency*. We can only be aware of inconsistencies in our belief or preference set when the contradiction is apparent to us—when both of the contradictory attitudes are "in the workspace" at the same time. Global coherence really should be viewed in the deflationary sense of coherence *within the global workspace* at a given time—it's not mind-wide coherence, just working memory-wide. But, as I have argued, there is no centralized, global executive that brings appropriate items into the workspace for processing. Rather, what's in the workspace at any given time is merely what has risen to the top of a massively parallel, unconscious, heuristically-driven recall and sifting operation, involving multiple layers of integration, associative priming, and biased prioritization of the outputs of certain systems over others. So, *sometimes,* inconsistencies can be brought out into the light, though often they won't be, and even when a *local* inconsistency is resolved (in one context), the same inconsistency may reappear in another

context, involving different subsystems, *precisely because there is no centralized global updating*. Local updating can take place, but in many cases, a resolved contradiction in one domain will leave the constituent parts that led to the contradiction in place (the compartmentalized pieces that contributed to the formation of contrary beliefs, for example). As a result, the ostensibly resolved contradiction may subsequently re-appear: the underlying modular integration and interface functions may simply put the pieces back together again, reconstructing the contradictory belief state, having "learned" nothing, globally speaking. If the *repeated* inconsistency is noticed, perhaps it can consciously and slowly be ameliorated (such as, for example, that point when I do *eventually* stop reaching for the light switch when I know the power is off).

A second way in which the sub-optimality of the system I am defending will reveal itself is in the inconsistent and non-ideal application of reasoning principles—the *cognitive reasoning biases* to which we are prone. For example, thanks to the *representativeness* and *availability* heuristics, we will be poor judges of probability, resulting in poor risk assessments; we will prioritize confirmatory evidence, and hence bootstrap up confidence levels for the ideas and beliefs we already possess, regardless of whether that confidence is objectively warranted; due to *anchoring and adjustment* we will form preferences relativistically, in ways that routinely violate transitivity, and are highly sensitive to arbitrary decoy options in choice sets; etc.

A third, more severe, way in which the sub-optimality of the system will reveal itself is in the persistence of *misinformation effects* and *perseverance of false belief*. This is for the same reasons that underlie inconsistency unawareness, just discussed, but it can be exacerbated by the fact that many motivational and action-guiding system may have a direct line to System 1 resources, and bypass conscious awareness for the most part. Since memories are encoded in parts, and the reconstruction of those parts is sensitive to shifts between the recall context and that of initial encoding, some "pieces" of memories may fail to be recalled, and the memory will hence be revised (unwittingly). If the memory is a memory of what we believe, or if we are evaluating a belief based on evidence we remember, this will cause all sorts of difficulties and be extremely error-prone. And certain aspects of recall contexts that are incongruent to aspects of encoding contexts will result in inhibition of some memories, which nevertheless continue to subsist in the system, waiting to be picked

up by another, different, recall cue. As a result, we should expect there to be occasions when, consciously, one disavows a particular belief, yet motivational and action-guiding systems may still be acting according to it, precisely because the disavowal stage is disconnected from the action-guiding and motor control systems: the contexts may be so different that the belief is recalled in one, but not the other. We will only notice the disconnect in certain situations—e.g., the *alief-like* scenarios discussed by Gendler. Most of the time, what we consciously avow *does* coincide with what our System 1 processes implicitly seem to "believe" (or alieve, or however you would like to describe it). And most of the time when the two don't coincide, it will be of so little consequence as to escape noticing. However, ongoing contradiction between avowed belief and behaviour can have deleterious personal and social consequences, which is presumably why our norms of rationality proscribe such states.

The *most* severe way in which the sub-optimality of the system will reveal itself is in cases of what we might call *pathological belief*—i.e., delusions and elaborate self-deceptions. I will argue that some monothematic delusional syndromes are the inevitable result of breakdowns or deficits at the level of secondary perceptual integration modules: the level at which cross-modal error-checking and perceptual smoothing takes place. The result, in some very specific cases, will be the generation of highly elaborate, content-rich false beliefs that are relatively immune to revision, as their formation is encapsulated in such a way that cuts off all alternate routes around them during revision. (I will elaborate on this argument in chapter 8).

In PART III of the dissertation, I wish to turn to empirical evidence from psychology and neuroscience to underscore the plausibility of the account I have given above. In chapter 7, we will look at memory as a direct analogue of belief. As I have argued, the computational challenges inherent in deliberation and belief revision *begin* with the tractability of memory searches. The limits of recall are the key to unlocking the frame problem. Whatever we can know for sure about how we remember, and how we forget, will tell us exactly what we need to know about how we *believe* and how we *unbelieve.* In chapter 8, I will look at pathological belief states and monothematic delusion as further proof of the belief revision account I have sketched, arguing that delusions are the systematic pattern of breakdown that reveals and confirms the modularity of belief revision.

Notes for chapter 6

[1] Note that some theorists suggest that whatever "workspace" we are talking about is likely to be associated with structures in the dorsolateral prefrontal cortex and anterior cingulate (Dehaene *et al.,* 1998; Sergent & Dehaene, 2004; Dehaene & Naccache, 2001; Dehaene, Sergent, & Changeux, 2003), though finding a particular "space" is difficult, as the whole idea of a global workspace is that it can draw in activations from all over. As a result, observed activations across brain structures could be indicators of localized "fetching" of information for the workspace, or that activation could *be* the workspace—i.e., the "space" is an emergent constellation of activity that isn't bringing fetched information *to* a particular locale. This would make more sense, generally, given that a global workspace in a particular area would be prone to catastrophic failure if that location were damaged. Obviously, certain key operations regarding the space may be localized in specific structures, but the space itself need not be a space.

[2] Or perhaps not—if the attentional system is constrained/limited in some way, it could still be plausibly tractable to think of it as a sort of spotlight trained on various items. For one, we know that attention is easily *hijacked* by unconscious (System 1) processes. As for "conscious" deliberate (System 2) uses of the attentional system, there is some evidence that attention is limited *even* when we are "consciously" paying attention: certain effects such as *inattentional blindness* (Simons & Chabris, 1999; Chabris *et al.,* 2011; Memmert, 2006; Mack & Rock, 1998) in which giving oneself a task to watch a video looking for *one thing* can cause one to completely miss another thing (a gorilla!) that would usually capture attention all on its own. Similarly, subjects have been shown to not notice the substitution of a person serving them, when one ducks beneath the counter momentarily and another person pops up—many do not notice the switch (Levin *et al.,* 2002). Another effect is *attentional blink* (Raymond *et al.,* 1992; Shapiro *et al.,* 1997) in which when attention is trained on the identification of a certain target, the moment a target is recognized there is a brief attentional "blink" in which a repeated instance of the target is missed. All of this points to an attention system, that at least once *engaged*, operates according to a fixed and impenetrable program. If you want to penetrate it, you have to re-engage it, or redirect it. The process that takes care of that redirection is obviously less conducive to a modular description, of course.

[3] 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).

[4] Additionally, there is evidence that young children engage in "syntactic bootstrapping"—generalizing from syntactic frames to derive meaning (and form concepts). Hirsh-Pasek & Golinkoff note studies by Gleitman & Gillette (1995) which seem to suggest that children are quite adept at identifying the meaning of verbs by attending to verb argument structure: "that children analyze events into predicate-argument structures, and that they link sentences to the event structure that they parse [...] the child inspects not only the world, but also the syntactic contexts in which a verb is used, to make predictions about its meaning" (Hirsh-Pasek & Golinkoff 1996: 125-126). Karmiloff & Karmiloff-Smith also note this type of early mapping of lexical constraints onto the world by children (Karmiloff & Karmiloff-Smith 2001: 71), and highlight studies by Gerken (1994) which show children are much better at remembering syntactically correct nonsense strings than ungrammatical ones of equal syllabic length (2001: 99).

[5] Carruthers (2006a) describes something very similar, using the language production faculty to pick up the outputs from *any* subprocess, formulate them (or decompose, or combine them), and "globally broadcast" them back for pickup and further processing by devices that "recognize" useful inputs. Although Carruthers clearly aligns his "global broadcasting" picture with Baars "global workspace" account, Jackendoff is loath 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). On this point, I am inclined to side with Jackendoff—the semantic priming idea seems much more plausible, especially if you conjoin that idea with many of the ideas regarding conceptual frames and memory organization that I have discussed earlier.

[6] This is to say, the *likelihood* gets updated with each iteration of the process: what we expect to see given the evidence *today* is a function of what we saw *yesterday*.

[7] I'll note in passing that this is reminiscent of the massively modular story Carruthers (2006a) wants to tell—letting abductive inference piggy-back on the recursive power of natural language production to keep "reframing" the question in an iterative attempt to re-query the modular sub-systems, for filtering. Carruthers suggests this involves cycles of "inner speech" broadcast out to the system to essentially "try out" through the subcomponent processors and report back. As mentioned in note 5 above, Jackendoff objects to the global "broadcast" idea, but in principle, the prospect of an iterative Bayesian evidence-checking algorithm which poses (and reposes) the question in finer form until a satisfactory "answer" resonates seems to be the best strategy available for programming a *machine* to engage in abductive inference, and hence may be the best strategy for explaining how *we* do it, without resorting to a sort of pessimistic Fodorian mysticism about it.

[8] Remember, his secondary search is after having soft-filtered down to 100 candidates: so the second search is confirmatory with respect to those 100—evidence will now be viewed *in light of them*.

[9] For more discussion of how Watson's algorithms differ from standard machine language and translation programs, see Bach (2011).

[10] From Searle's *Chinese Room* (1980) on down. *Cf*., Dreyfus (1972; 1986; 1992) for a criticism of computationalist analogies of human intelligence in general.

[11] From personal experience teaching a class about this study, I have had to draw Venn diagrams to illustrate exactly *why* it cannot, logically, be true that 'feminist bank teller' is more likely than 'bank teller'—and *even then*, I have students who argue the point! System 2 can be *very* slow at times.

[12] If you thought "Pop!" there's your availability heuristic at work. Similar effects are used in children's riddles, such as: "A boy and his father are in a car accident, and the father is killed. When the boy arrives at the hospital, the head surgeon says 'I can't operate on this boy – he is my son!' How is this possible?" The obvious correct answer is often *unavailable*, due to the prevalent stereotype that a "head surgeon" would not be a woman. This is an example of the confluence of availability and representativeness.

[13] Median per capita income, worldwide is only $1225/year. The line that marks entry into the top 1% is only $34,000 (Milanovic, 2014).

[14] I think this is a highly under-appreciated point that should bother rational choice and preference valuation theorists more than it does: in Ariely's study, the presence of a fully dominated option *reverses* preferences regarding the items on the choice list. A dominated option should be entirely *inert*—and it is, in the sense that *no one considers it*—but its inclusion has a mediating effect. So a dominated option weirdly dominates a preference valuation. This is a point I want to return to when we discuss the entrenchment of "irrational" delusional belief in chapter 8—delusional beliefs, too, are often behaviorally and/or cognitively inert, and yet they may well affect other cognitive activities in dramatic *indirect* ways—just as the dominated, ostensibly inert, option does in Ariely's choice set.

[15] Of course, I didn't cancel it after 3 months. When I didn't have satellite radio, I didn't want it enough to pay $15 a month. Once I have it (for free at first), my anchor shifts, and now it's suddenly worth it. This is an example of *loss aversion*, the *endowment effect,* and the *status quo bias* (Kahneman, Knetsch, Thaler, 1991).

[16] This confidence that we all tend to have that *we* wouldn't fall for it is evidence of yet another bias at work: the *optimism* or *self-confidence bias* (Gilovich, 1991).

[17] McKay & Dennett (2009) make a different, but quite interesting argument for "the evolution of misbelief" that misbeliefs serve a valuable purpose as "doxastic shear pins":

> We envision doxastic shear pins as components of belief evaluation machinery that are "designed" to break in situations of extreme psychological stress (analogous to the mechanical overload that breaks a shear pin or the power surge that blows a fuse). Perhaps the normal function (both normatively and statistically construed) of such components would be to constrain the influence of motivational processes on belief formation. Breakage of such components therefore, might permit the formation and maintenance of comforting misbeliefs – beliefs that would ordinarily be rejected as ungrounded, but that would facilitate the negotiation of overwhelming circumstances (perhaps by enabling the management of

> powerful negative emotions) and that would thus be *adaptive* in such extraordinary circumstances. (2009: 22)

Cf. Dretske (1986) for a very different and thorough examination of how best to model and explain how mental systems *misrepresent* information, and how to count (and account for) misbelief.

[18] Aside from a "need" to be at par, resource-wise, with my social group.

[19] I should note that Samuels nevertheless essentially agrees with Fodor in the end, when it comes to central systems: "the most plausible position to adopt is one that takes a middle way between those, such as Carruthers, who endorse a thoroughgoing massive modularity, and those, such as Prinz, who reject modularity altogether. The situation is, in other words, much as Fodor advocated over two decades ago" (Samuels, 2006: 52).

[20] Note there are some tool purists who would argue that if you could only have *one* sort of tool, it should be a *hatchet* (Paulson, 1988). Perhaps I could rework my analogy. But for now I'll stick with hammer.

[21] I am not arguing for the existence of all or any of these—they seem plausible enough to me, but nothing in my argument hinges on whether these particular modular assemblies work out.

PART III

7 Memory distortion

In this chapter, I want to look specifically at current empirical research into memory, specifically with an eye to studies examining how we *store, retrieve,* and *revise* memory, in order to seek empirical confirmation of some of the four predictions that fall out of my modular account of belief revision, highlighted at the end of chapter 6. I suggested 4 specific sorts of sub-optimal functioning through which in which a modular, heuristically driven system of belief revision would reveal itself:

(1) *pervasive unawareness of inconsistency*, due to processing limitations of the amount of data under consideration at a given time;

(2) *cognitive reasoning biases* that would reveal themselves in complex deliberative settings that our systems were not adapted for;

(3) *behavioral, attitudinal, and cognitive effects of false beliefs* that have gone undetected in the system as a result of (1) and (2);

(4) the *generation of certain classes of belief that are incorrigible or irremediable*— beliefs which whose generation is inevitable given the functioning of sub-serving modular processes, and which may be (ironically) regenerated and solidified via revision.

In this chapter on memory distortion, I will discuss a host of experimental findings regarding the limitations of remembering and forgetting, which both confirm the above predictions, lending support to the model I have defended, as well as elaborate on and illuminate how that model works in practice.

Before jumping in to the discussion of the empirical literature, I should note that there is an interesting elision in the psychological literature from *memory* to *belief*. Psychologists tend to move back forth between these two terms as if they are generally interchangeable, or at least that the progress from memory to belief is essential or inevitable (i.e., distort a person's memory, they will end up with distorted beliefs). For example, studies on the *misinformation effect*, some of which we mentioned in the introductory chapters to this

thesis, and which we will examine in more detail below, cite and refer to each other constantly despite the fact that some are explicitly titled and described as studies on *false memory*, whereas others are titled with reference to *belief perseverance.* The *misinformation effect*, as we shall see, is an effect of *memory distortion*, which reveals itself in *false belief.*

In general, I expect philosophers would likely prefer more precise definitional boundaries the loose interchangeable usage of "memory" and "belief" in the psychological literature, but I would argue it's actually potentially helpful for our discussion. I think that a close analysis of the common findings between these studies on memory, forgetting, misinformation and belief perseverance will help us tremendously to sort out clarify and in many ways confirm my account of belief revision. I have already argued that everything we want to say about belief acquisition and revision depends on what we can say about memory, insofar as all deliberation, all comparative evaluation, and all evidence checking are mediated by (and begin with) a recall step, and hence inherit all the limitations and computational tractability constraints inherent in recall. Every psychological or empirical claim about belief revision and its limits is extensionally equivalent to a claim about the limits of memory revision and recall. And as we have seen, in the discussion of Tulving's account of memory, the limits of recall both support the sort of massively parallel modular organization of cognitive functions that I have argued for, as well as help explain how *seemingly* global cognitive operations can approximate Quinean ideals of deliberation and belief revision without actually being Quinean. Below, I will look in detail at some of the empirical findings from psychology regarding memory distortion. I contend that studies of memory—specifically studies of how, and under what circumstances we can revise memories, and *forget* (former) memories—lend support to the sort of belief maintenance system I have defended above, predicated on massively parallel assemblies of modular systems with built-in limitations and heuristically driven mechanisms for evidence-sifting, relevance determination, inferential reasoning and judgment. The manipulation conditions of memory are the systematic pattern of breakdown we should expect from modular systems of the sort I have defended in PART II. First I will highlight and give on overview of the research findings, and then I will explain in what way these findings confirm the predictions that I laid out in the end of chapter 6.

7.1 Manipulating memory

In this section, I will look at a number of studies highlighting the various ways in which memories can be manipulated in experimental settings—both implicitly, so as to (unconsciously) bias subsequent reasoning; and explicitly, insofar as rich false memories can be instilled in subjects, who subsequently profess to believe with high degrees of confidence that their memories are veridical. The purpose of highlighting these studies is to try and delineate the specific conditions under which misinformation, false memory, and false belief can gain a foothold and proliferate within the cognitive system, and in what ways it can influence judgment and behaviour (i.e., predictions 1 and 3, above). I will attempt to show that the conditions enumerated will support the idea that memory systems encode, store, tag, retrieve and (sometimes) revise memories in ways that suggest modularity at every stage—encapsulated processing, automatic activation, and heuristic-based organizational strategies. The ways in which memory can be experimentally manipulated will be shown to be analogous to the way perceptual illusions can manipulate phenomenological reports of perception. And, given the direct analogy to perceptual illusions, I will argue that just as perceptual illusions are the telltale calling card of modular systems, so too are the illusions and distortions of memory and recall the telltale symptoms of modular systems of belief.

7.1.1 The misinformation effect

The *misinformation effect* refers to phenomena where the content and/or retrievability of memories can be corrupted or distorted by the presentation of or exposure to misleading post-event information. A few cases of this effect were mentioned back in the first chapter, as cited by Harman in her discussion of belief revision failures. Recall Anderson & Ross' (1975) study, in which subjects continued to make judgments based on misinformation, even though they had been fully debriefed that the information was false—other beliefs and judgments implied by the false belief persevered regardless, and even *further ones* were still made based on the information.[1] As Anderson & Ross note in a follow-up study (1980):

> a theory concerning the relationship between two variables—generated through exposure to a minimal data set—can survive even a complete refutation of the formative evidence on which the theory was initially based. (Anderson & Ross, 1980: 1043)

Wilkes & Leatherbarrow (1988) found that study participants who generated inferences based on misinformation, were just as likely to maintain those inferences after debriefing as subjects who had *never* been debriefed—"people failed to edit elaborative inferences made during reading before the correction occurred... not even an explicit and direct denial was sufficient to purge the memory record of all of its implications" (Wilkes & Leatherbarrow, 1988: 378). Wilkes & Leatherbarrow conjecture that there are likely two distinct memories constructed:

> One way of reconciling this discrepancy between free recall and comprehension is to use the distinction introduced earlier between a memory record for the text content and a different record for an associated situation model. The text base encodes the literal content of a message sequence, and the situation model acts as a representation of what the content means when it is interpreted by means of related knowledge schemas in memory. (1988: 379)

The idea here would be that subjects remember the initial information as text, and also separately construct a "situation model" or narrative contextual memory that incorporates that text (interprets it, puts it to work). When misinformation is introduced, the same sorts of processes ensue, and a new, contradictory situation model will be constructed—the misinformation doesn't necessarily overwrite or displace the correct information, it is stored in parallel. The *problem* comes when subsequently free recall is employed and a "choice" must be made—Wilkes & Leatherbarrow argue that simple *availability* and *recency* favor the most recent construction, in this case the one built on misinformation. Later, after debriefing, a subject may have been led to re-tag the false information as false, but unless *specific* attention is paid to updating the misinformation-mediated situation model (and any further inferences or explanations that were based on it), then, as it is stored separately from the text content on which it was based, it will survive the debriefing.

> Where two related but independent episodic records stand in contradiction to each other, free recall is biased towards reproducing the content of the more recently presented version, whereas this does not occur if the records are consistent with each other. A text base with such tagging can then be used to generate new inferences as required and for locating where old inferences persist. On the present data it seems that editing relies more upon using the most recent version of contradictory evidence to generate new inferences than it does on the old version for locating errors in the record. It is too early, however, to conclude that this function never applies. Deciding what entries in the complex record will qualify for change once a contradiction has been accepted cannot be straightforward, as they include both the consequences of the new information that need to be added and the consequences deriving from the old information that need to be discounted, and these sets need not be exact inverses of each other. (Wilkes & Leatherbarrow, 1988: 380).

In a related study, Wilkes & Reynolds (1999) established clear evidence that the *fruitfulness* of misinformation plays a large part in how easily it can be entrenched into memory and belief. (Mis)information that can immediately be put to work in causal explanations, or the generation of plausible elaborative inferences, will be more firmly established. Worse, this will result in simultaneous inhibition of the correct information that it has displaced—a finding similarly supported by Ross, Lepper & Hubbard (1975), Anderson, Lepper & Ross (1980), Schul & Burnstein (1985), and Seifert (2012). There are also studies that have established that *repetition* of information, regardless of whether there is any supporting evidence, will result in higher levels of retention and endorsement. As mentioned in the previous section, simply entertaining a proposition gives increases its perceived validity (Hasher, Goldstein, Toppino, 1977). This supports the idea we looked at in chapter 1 from Gilbert (1991)—the 'Spinozan' system, in which all propositions must first be believed, and only then may they be subsequently evaluated in a second step. This, of course, requires *first* encoding the information into memory, and then *rehearsing* it in order to evaluate it.

Anderson & Bower (1973) report that associative links made from false information believed at one point, and then subsequently (explicitly) disbelieved, will persevere unless they are *positively* undermined with competing associations that can take their place. Similarly, Nisbett & Ross (1980) established through a number of experiments that the debriefing *can be* successful only when the phenomenon of *belief perseverance* is made salient to subjects *and* if the false information is positively undermined in such a way that all associations, causal explanations and implications that flowed from the initial misinformation are also specifically brought to attention *and replaced with new ones*. Johnson & Seifert (1999) note that powerful causal explanations make good replacements regardless of likelihood—a good story takes precedence. Again, here we seem to see the workings of our heuristics and biases—*availability* and *representativeness* in particular exert a strong hand insofar as revision and updating go. This supports the model I have described: what we see going on here isn't belief revision so much as belief *displacement.* We preferentially recall information based on heuristic search algorithms and associative processes that that privilege certain items (recent, available, causally structured, those activated in associated contexts) over others. But if it is the false or incorrect information that is more recent, available, built

into causal structures, or co-activated in associated contexts, then *it* will be what we recall when we inspect our beliefs.

Many of these studies work with a *jury* paradigm—which has real world significance given the very real possibility that prejudicial information that jurors may have heard (and believed, at least provisionally) could affect their judgment, even if they were told to disregard it. Numerous studies have highlighted the practical difficulties in getting juries to disregard information—we have seen numerous misinformation studies already that use a *jury* paradigm, and all show consistent results that once information is presented, it *can't* be disregarded completely. Even when the "jury" members are consciously aware that the information is false or irrelevant, *it still colors their judgment*. Recall Gilbert, Krull, & Malone's (1990) study, discussed above in chapter 1, which used a jury paradigm, in which subjects were given crime reports in which some information had been color-coded and marked "untrue/disregard". The result was clear: "the number of false statements that subjects misremembered as true was reliably correlated with the length of the prison term they recommended" (Gilbert *et al.,* 1993). This last point is one that has serious practical significance, given the numerous points in any trial where jurors may be asked to "disregard" something they just heard. Evidence suggests that the disregard instruction, even in cases where it is understood fully, and where jurors consciously believe they have successfully disregarded as instructed, fails to block the false information from continuing to infect and motivate judgment.[2] This is known as a "boomerang effect", where "the admonishment draws added attention to the information in dispute, heightening its salience and accessibility in memory" (Kassin & Sommers, 1997: 1047). Worse, the more sternly the disregard order is given, the *less* likely jurors are to heed it (Wolf & Montgomery, 1977). Pickel (1995) suggests that "the backfire effect may not occur because jurors *deliberately* disobey the judge" but rather that merely entertaining the proposition reinforces it.[3] This, again, is evidence of associative processing which is not (and likely cannot) be mitigated by global coherence checking. Repeated activations make certain items *more* retrievable, and hence more likely to be recalled consciously, and more likely to be conscripted into the purely associative (and unconscious, subdoxastic) work of System 1 processes. Wistrich, Guthrie & Rachlinski (2005) even found that *judges* were fairly limited in the degree to which they could ignore evidence they themselves had deemed inadmissible: by telling themselves to

ignore something, they inadvertently primed themselves to recall it. This seems to be a pretty clear indictment of the Quineanism of belief revision.

7.1.2 Rich false memory

Elizabeth Loftus and her colleagues have done numerous studies on the *misinformation effect*, and the ease with which memory can be manipulated, and *false memories*—even richly detailed false memories—can get a foothold in one's belief system. This research obviously has huge practical implications, especially when it comes to the role of memory and testimony in courts of law. Some of her initial work shows how effectively post-event misinformation seems to literally *overwrite* episodic memories, in such a way that the misinformation is incorporated completely into the memory of the event, and cannot be reversed. Loftus' (1974) studies used automobile crash incident reports, and found that subjects' subjective descriptions from memory could be manipulated by various post-event methods. For example, subjects shown films of car crashes would estimate higher speeds for the cars involved depending on the wording of the prompt—if asked "how fast would you say the red car was travelling when it *impacted* the blue car?" subjects will generally report lower speeds than if asked "how fast would you say the red car was travelling when it *smashed into* the blue car?" Memories of pictorial images could also be manipulated. Loftus *et al.* (1978) demonstrated that subjects shown pictures of an accident scene with a clearly visible "yield" sign, were prompted to remember falsely that they saw a "stop" sign if the subsequent description of the accident mentioned a stop sign. Eyewitness face identification tasks are also easily manipulated. Police lineups and suspect picture arrays are notoriously bad for this: if a subject has reason to believe that the actual suspect is pictured (or in the line up), but the suspect is *not* in fact present, it's possible that the face of an innocent, but similar looking person can *overwrite* the eyewitness memory, such that even when faced with the original, actual suspect, the subject will recall the substitute as having been the culprit.

Some of Loftus' most amazing findings, however, are regarding the ability of experimenters to instill rich, detailed false memories in subjects—memories that even after debriefing, even when proven to be *impossible,* subjects will insist they actually remember. In one series of studies—using what Loftus (1997) calls the "lost-in-the-mall" paradigm—subjects were prompted to remember dramatic childhood events that never actually occurred

(such as being lost in the mall, or having been hospitalized overnight), based on a false narrative credited to a family member. For instance, subjects are told "your mother told us about your hospitalization when you were seven..." and many subjects suddenly develop and elaborate on this "memory". Porter, Yuille, & Lehman (1999), using Loftus' design, were able to plant a false memory of a childhood animal attack in half of their subjects. In the most dramatic study, Braun-Latour, Pickrell & Loftus (2004) managed to get participants to form elaborated false memories of having met Bugs Bunny at Disneyland—though that clearly isn't possible, as Bugs is a Warner Bros. character. Many subjects had highly specific memories of having gotten Bugs' autograph, or having touched his ears, etc. In the Bugs study, a verbal suggestion alone (i.e., "Did you meet Bugs Bunny?") elicited false memories in 17%, but having a *picture* of Bugs Bunny on the questionnaire, and no mention of him in the question, elicited a 48% hit rate for false memory (2004: 14).[4] The fact that the pictorial suggestion was more effective than the verbal cue is interesting, and supported by Schacter *et al.* (1996) who also elicited higher recall rates via pictorial representations than words. One of the conclusions Loftus draws from these studies is that *subtlety* matters—a subtle suggestion of misinformation is more likely to make it into memory. Tousignant, Hall, & Loftus (1986) note that recollection in many false memory and misinformation studies is enhanced by subtlety of presentation. A greater effect will be elicited "by the misleading object of an auxiliary clause than by the same info in the focus of the question – e.g., 'did the intruder *with the moustache* say anything to the professor?'" (Tousignant *et al.,* 1986: 330).

One possibility that we might consider is that there is *not* a general lesson about memory manipulation to be drawn from these studies, but rather, some people just are very prone to it—either due to excess suggestibility, or to impaired memory, or tendency to confabulation. Loftus (2005) argues that this isn't the case. For one, subjects who report false memories don't perform below norms on standard memory and recall tests, which suggests no special general memory impairment is at work in those people. Secondly, in a variation of the misinformation memory-planting experiment, McClosky & Zaragoza (1985) found that when the misinformation item is *excluded* from the response choice set, subjects choose the *correct* information. They show participants pictures of a crime scene in which a hammer is clearly present; later they attempt to plant the misinformation that it was a screwdriver at the scene. When asked to respond as to the tool, many will choose

screwdriver over hammer, given those choices. But if asked to choose between hammer and a *novel* item (wrench), in the absence of the (misinformation) screwdriver option, they choose hammer. This suggests that the original memory is actually there and intact, but is inhibited by the activation of the "replacement" information (screwdriver). Without that inhibition, subjects resort to the veridical memory. The upshot of this: false memory implantation is not a sign of a faulty memory in general; it happens to people with apparently fine memory.[5] Note here that again we see evidence not of memory revision, per se, but something more like memory displacement—just as I argued above with respect to belief displacement. Note that this goes a long way toward explaining pervasive inconsistency unawareness in individuals: contradictory beliefs can displace one another, but various contexts may employ *one* or *the other*, without co-activating the two and making the contradiction apparent.

The key, according to Loftus, that distinguishes those who manage to *resist* false memory is mostly circumstantial: there appears to be a small window during the misinformation induction stage in which what Tousignant *et al.* (1986) refer to as "discrepancy detection" can take place. If the information is detected at the time of initial exposure—if the inconsistency with previous memory is present before awareness at that time—then the subject can resist, or at least look for clarification, more evidence etc. Once this critical window has passed, however, it can be very difficult to retroactively detect the discrepancy.

7.1.3 Discrepancy detection

Recall that, according to Loftus, one of the conditions favorable to false memory implantation was *subtlety* of presentation—e.g., the *picture* of Bugs on the Disney pamphlet is more likely to elicit the memory of having met him than a question asking directly; putting misinformation in an auxiliary clause makes it more likely to stick, etc. The fact that subtlety works so well highlights the fragility of the *discrepancy detection* window. In all the cases where false memory is successfully planted in these experiments, subjects have, within their cognitive possession at the moment when presented with the false information, *other* information that immediately contradicts the suggestion. This should (ideally) be recognized as an inconsistency that demands attention, and presumably the false information would be

blocked. If the false information or suggestion, upon presentation, *activates* the memory or previously held belief that it contravenes (e.g., I was never bitten by a dog; Bugs wouldn't be at Disneyland, etc.), then the suggestion will be properly evaluated. But if the contravening evidence in memory is not activated, then the discrepancy goes undetected, and the false memory is taken as veridical. Subtlety of presentation merely decreases the likelihood that the suggestion will activate associated memories contrary to it.

Recall Cherniak's point about the failure of normative accounts of rationality to respect the long-term/short-term memory distinction:[6] he suggested that one cannot evaluate inconsistencies unless *both* of the beliefs in the inconsistent set are brought into short-term memory and hence awareness synchronously. Discrepancy detection requires exactly this—something has to trigger the previously held belief (or memory) to a degree sufficient to bring it to consciousness for comparison and evaluation. Subtlety of misinformation presentation reduces the chances that the original, inconsistent, memory or belief will be activated, and the opportunity for discrepancy detection is lost. Worse, later instances of recall will favour the most recent, which happens to be the false one. Discrepancy detection is not only a short window, it may in many cases be a one-time window, never to return. So in this way, coherence checking in general is subject to the limits of recall, and will be heavily circumscribed by the transience of the discrepancy detection window.

What other conditions, aside from subtlety, affect the ease with which false memories can be planted? Braun-Latour *et al.* (2004) suggest a number of variables that have a positive affect on false memory implantation, some of which might not be surprising at all:

- The more similar the suggestion is to actual events, the greater the effect.
- The more credible the source, the greater the effect.
- The more plausible the information, the greater the effect.
- The more frequently exposed the information, the greater the effect.

The first three conditions on this list seem obvious: clearly, credibility of source, plausibility of information, and similarity to actuality are all likely to decrease discrepancy detection, as they reduce the discrepancy. The interesting item on this list is *frequency*—that seems counter-intuitive at first. Shouldn't *repetition* of false information or a false suggestion *increase* the opportunity to detect the discrepancy? Perhaps, though as we noted above citing Hasher & Zacks (1984): frequency of occurrence is apparently encoded separately from other

features of event being remembered, and this frequency counts as a point in favour of a particular item in any subsequent evaluation—so the increased opportunity for discrepancy detection is working at cross-purposes to the automated process that "scores" information based on frequency of occurrence in the case of misinformation.

Roediger, Jacoby & McDermott (1996) argue that repetition of misinformation *retrieval* is actually what increases the likelihood that the discrepancy remains undetected. In their study, Roediger *et al.* found that once the false information gets a foothold—i.e., is not detected on first pass—then increased exposure to it will activate the stored false memory, rehearsing it, essentially, and thus strengthening it. They chalk this up to the "testing effect" that has been known about memory presumably since people started trying to remember things: simply that "the act of recalling or recognizing material generally increases the likelihood of its later recall or recognition" (1996: 2). Roediger *et al.* also suggest that repeated retrieval of the false info has a double effect: not only does it rehearse the false information, it also has an inhibitory effect of any contradictory information in memory (which in this case is correct).[7] This, again, is precisely as predicted by the account laid out in PART II of this dissertation.

7.2 Intentional forgetting

Given the challenges associated with misinformation effects, false memory and belief perseverance, it would be useful to understand the mechanisms by which these challenges *could* be attenuated. Under what conditions can misinformation effects and false memory be *blocked?* Under what conditions can a subject be *disabused* of perseverant beliefs? If we are aiming to have an account of belief revision that respects the limitations of human cognition, we will need to firmly delineate the boundary conditions for these memory corruptions and *de*-corruptions. One research program that might give us valuable insight in this regard involves *directed* or *intentional forgetting*—the study of how, under what conditions, and with what limitations are we able to *un*-remember previously stored information. As we shall see, there are only a very few circumstances in which *actual forgetting*—including the wiping out of both explicit and *implicit* memory traces—can happen. The reason for this, as we shall see, is that the "forgettability" of items is essentially baked into the manner in which the initial memory was encoded. I have already suggested that this plays a large role in our

belief revision "failures": we don't revise belief as much as displace it. But the original belief is never fully forgotten, and could still be recalled in certain contexts, or could be reconstructed out of the pieces from which it formed initially, or could be conscripted into the working of subdoxastic System 1 processes. The point is, we won't notice the lingering effects of the "revised" (displaced) belief unless we co-activate it with the one is was ostensibly replaced with—and given they are contrary, the context priming of recall will often lead to the inhibition of one as a result of the recall of the other. Our brains are set up to make it very easy to remember, and very hard to forget. Indeed, the best strategy to avoid misinformation effects, false memory and attendant perseverant false beliefs is to avoid remembering in the first place—but that will be very difficult to do, given that that very act of consciously deciding to *not remember p* will cause you remember *p* (as *p* is embedded in the memory you have of deciding not to remember *p....)*. Let's look at the research to see what it teaches us.

7.2.1 The DF experimental paradigm

The *directed forgetting* (DF) experimental paradigm is quite simple, and the findings are robust. A standard experimental set-up would be as follows, taken from one of the first such studies, Bjork & Woodward (1973):

> 40 undergraduates were presented seven 24-word lists, each consisting of a random mixture of 12 R[emember] and 12 F[orget] words; the cues to remember or forget were presented subsequent to each word in turn. Six lists were followed by either an immediate test of R-word recall or a distractor activity, and 1 list was followed by a prearranged signal to recall both R and F words from that list. (1973: 22)

The results of this and similar experiments—where subjects are instructed after each word-item to "remember" or "forget" the item—show that the F-cued items (the ones subjects were told to forget) are nevertheless often remembered. On explicit memory recall tests, subjects generally *do* recall the R-cued items better than the F-cued items, which suggests successful forgetting. However, on *implicit* memory tests, studies have repeatedly found that subjects just as readily "remember" F-cued words as R-cued ones.[8] Implicit memory is tested for using stem-completion tasks—subjects are given word-stems corresponding to both the F-cued and R-cued words, and successfully complete *both* with the same level of accuracy. (Note that the stems given are ambiguous enough that they could be *lots* of different words—

e.g., if subjects were presented an F-cue for *STRAIN*, on the stem completion task they might be given ___*AIN* or *ST*____. Those stems when given to control subjects (who were never given any words initially) would not be statistically likely to come up with *STRAIN*.) The key finding here is the distinction between explicit recall and implicit memory: given a free recall test, when asked to recall *all* the words, F-cued and R-cued alike, subjects tend to remember the R-cued ones at a much higher rate, "recalling" very few of the F-cued words. However, on the stem-completion task, subjects correctly complete the stems at roughly the same rate for F-cued and R-cued words, which suggests they are equally well "remembered" implicitly. Paller (1990) found implicit effects of F-cued information weeks later.

One explanation of this is that memory is stored separately and redundantly for implicit vs. explicit retrieval. Wang (2010) reports that in studies of patients under anesthesia or sedation, information can be presented that is remembered implicitly with no explicit recall at all. He cites Munich *et al.* (1993) who read *Robinson Crusoe* to surgical patients under general anesthesia, and then asked those patients, upon waking, to free-associate the word "Friday"—many talked of desert islands, etc., though none had any explicit recall of having been read a story (Wang, 2010: 178-79). Wang concludes that

> the most favorable conditions in which priming can take place, is where there is a high state of consciousness but in the presence of drugs which impair or disrupt explicit encoding. Whilst it may be of academic and experimental interest to the cognitive psychologist that such implicit memory effects can occur, there may also be important clinical implications relevant to models of the genesis of psychopathology such as post-traumatic stress disorder, other anxiety states, and sleep disorders. In sum, it may be possible to experience distress and psychological trauma in these circumstances of impaired explicit encoding, and subsequently be left with psychological disturbances, the origins for which the sufferer has no explicit recall. (181)

Further evidence for the separate encoding of implicit vs. explicit memory comes from studies of amnesiacs, who exhibit the "Warrington-Weiskrantz effect" (1974; 1978)—impaired performance on explicit memory tests (as one would expect), yet little impairment on implicit tests, such as stem-completions. Paller notes that:

> In normal subjects, one factor that like amnesia influences performance in explicit memory tests but does not influence priming is the level of processing during encoding. Priming measures in several different tests were not higher for semantically processed words than for phonemically or orthographically processed words (e.g. Carroll, Byrne, Kirsner, 1985; Graf & Mandler, 1984; Jacoby and Dallas, 1981). Similarly, priming levels were not changed by manipulations of elaborative processing during acquisition, of intention to learn, or of rehearsal duration. On the other hand, all of these encoding manipulations had robust effects on recall and recognition. (Paller, 1990: 1021).

There is something interesting going on here in the encoding stage: information is getting *in* at a level accessible for subsequent implicit retrieval—i.e., priming effects such as subsequent successful stem-completion, and judgment biasing that utilizes the implicitly remembered items. But that information is *not* being encoded in such a way that allows for successful explicit retrieval—i.e., free or conscious recall. And importantly, manipulation at the encoding stage can variously affect explicit recall ability, but implicit recall seems fairly immune to variation or manipulation at encoding.[9]

Distinct encoding routines for memories to be accessed implicitly and explicitly make a lot of sense with a lot of the ideas under discussion throughout this dissertation. For one, it supports the dual-systems, dual-process views: perhaps it is as simple as System 1 makes use of implicit memory processes, whereas System 2 can only operate under explicit recall conditions.[10] My supposition would be that encoding memories for explicit recallability necessarily involves making them available to conscious awareness—i.e., tagging them in some way that makes them "findable" to a conscious search. But, as we have seen argued repeatedly with respect to tractability, the very *last* thing we want is for conscious awareness to roam freely over the cognitive system. And we certainly don't want to be aware of every remembered item that is corralled into some form or other of processing by various sub-systems. Hence, it's plausible that all information that is stored, is stored redundantly: it would be stored in one iteration for implicit use—for future access and retrieval by unconscious systems via the global workspace—and may or may not *additionally* be stored in a way (directly) accessible to consciousness or recallable to the global workspace. Information that we wish to *forget* could in principle be removed from conscious-access storage (respecting the constraints mentioned above, of course), but that will leave the implicit-access "copy" of the information intact, and unconscious, subdoxastic subsystems may still be making use of it (they may have "hotline" connections to it that bypass conscious mediation). This will leave us (perhaps often) in a state where we *act* like we believe *p*, while we *think* that we do not believe *p.*

My supposition about separate encoding for implicit and explicit retrievability also fits with views we looked at from both Cherniak and Barsalou regarding how memories and concepts are stored and organized. Recall that Cherniak argues for a "compartmentalization" strategy that can expedite retrieval across domains. Barsalou similarly suggests that a

redundant storage system, despite using up more cognitive space, would have processing payoffs over a conservative storage system in which concepts inherit features based on type-relations.[11] Again, the idea would be that pretty much *everything* we come across that passes some threshold for storability gets stored at an implicit level, which means that it can be "recruited" by various subcomponent, modular systems that are not connected to awareness directly. *Explicitly* recallable information would have to be much more elaborately tagged in storage—as it would be stored as part of an episode associated with numerous other pieces of information.

7.2.2 When *can*—and when *can't*—we forget?

So, we have seen above that directed forgetting is achievable pretty much only at the explicit level. Implicit memories persevere—and this fact goes a long way to explain a number of the findings we have looked at so far in this dissertation, as well as lend support to some of the account we have examined regarding dual-process cognition. So, in this sense, we actually can't really *ever* fully intentionally forget. But what can we say about the circumstances under which we *can* and *cannot* successfully be directed to forget *explicitly*? Given that explicit recall *is* blocked in some cases by instructions to forget, what do we know about the circumstances under which that is more or less likely to occur?

Well, from the DF studies, we do have evidence that the *list* method is more effective than the *item* method at inducing forgetting (MacLeod, 1989; Basden *et al.,* 1993). In the *list* method, a number of items are presented and then a single instruction to "forget all the words so far" is given. This is more effective than the *item* method where words are given with individual cue to remember or forget each one. Johnson (1998) similarly shows that *global* "forget" cues are much more effective than *specific* ones. One possible reason for this is that the *specific* F-cue may accidentally reinforce the item in question (Schul & Burnstein, 1985; Wyer & Budesheim, 1987; Wyer & Unverzagt, 1985). For example, the subject is shown an item *p*, and then instructed to "forget" it. There is an implied anaphoric reference in the specific F-cue, that essentially *repeats* the item in question—in other words, what is happening is the subject attends to *p*, and then attends to the instruction to forget *[p]*. So the item gets a rehearsal, in a sense, merely by virtue of being the directly implied object of the F-cue. In the *list* method on the other hand, the subject is presented with *m, n, o, p, q, r, s, ...*

and then the global F-cue to "forget all of the items on the list". This is potentially one fewer rehearsal of *p*, for example, and hence less likelihood that it is maintained in memory.[12]

Another theory that could help explain the differential abilities to follow item- or list-oriented F-cues is so-called *ironic process theory* (Wegner, 1994; 1997; Wegner *et al.* 1990).

> Mental control is accomplished, in this view, by the interaction of two processes—an *intentional operating process* that is conscious, effortful, and interruptible and an *ironic monitoring process* that is unconscious, less effortful, and uninterruptible. The operating process promotes the intended mental control by searching for mental contents consistent with the intended state of mind, so, for example, this process might look for distractors when the person is trying to suppress a thought, or for signs of fatigue when the person is trying to go to sleep. The monitoring process, in turn, searches for mental contents signaling a failure to create the intended state of mind. In the case of thought suppression, for instance, the monitor looks for the to-be-suppressed thought. In the case of trying to sleep, the monitor looks for signs of wakefulness. The two processes function together as a feedback unit to produce mental control... The irony of the monitor, however, is that in providing the needed search for the failure of mental control, it increases the accessibility of exactly the most undesirable thoughts... As long as the operating process is healthy and unimpaired, this is only a small problem. The operating process if far more effective than the monitor given the luxury of the processing capacity it consumes, and so it usually overwhelms the slight sensitivity to counterintentional mental contents produced by the monitor. However, when mental load arises—in the form of distractions, stress, pressure, alcohol intoxication, or other impairment of processing efficiency—the operating process may be overtaken by the monitoring process in its ability to fill consciousness with the products of its search. Mental control then not only ceases, but works against itself. (Wegner, 1997: 48)

The upshot of this idea is that active suppression can serve to intensify persistence of memory (or by extension, belief).

Why would ironic processing happen, one might ask? One explanation is that this is a way to ensure at least some basic sort of *discrepancy detection* takes place. As we have seen, if we are not actively entertaining contrary evidence, or actively engaged in discrepancy detection, pretty much anything under consideration is capable of slipping in to memory, and extensionally, to belief. If activations generally provoke ironic activations, and hence bring contradictions to mind—literally, in working memory—then this gives some opportunity for evaluation. The cost of it, however, is that it also gives an opportunity for *rehearsal* of ideas or propositions that would ideally not be rehearsed. Additionally, recall the Roediger *et al.* (1996) study cited in § 7.1.3, above: there we saw that retrieval and rehearsal of misinformation, triggered by repeated presentation of it, actually *inhibited* the correct information from being recalled (explicitly). If we look at this with *ironic processing* in mind, we see that what may be happening is that the rehearsal of the misinformation *does*

activate the countervailing information implicitly stored memory, and an *implicit* discrepancy detection check will notice the discrepancy, and *mark the "old" (i.e., correct) information as false*—as the misinformation benefits from repeated exposure and recency. This, as Wegner notes, can happen easily when the retrieval context is stressed by other situational or affective pressures.

Of course, the list vs. item presentation mode of F-cues may not be totally relevant in figuring out how to forget things (misinformation, false beliefs, etc.) in the actual world, as we tend to engage *items* (episodes, facts) more often than *lists*—the list paradigm is something that really happens only in artificial settings. So the fact that F-cuing by list is more successfully than by item may not extend very far, in practical terms. The corollary finding (Johnson 1994) that *global* F-cues are more effective than *specific* ones might be more amenable to the real world. For example, if we want to disabuse someone of some misinformation they picked up reading conspiracy websites, we might instruct them that *everything on that site is false, disregard all of it.* Of course, outside of a small number of cases of egregiously bad information sources, it's probably not often *true* that *everything* from a certain source is wrong, or to be forgotten. And in day-to-day reasoning and evidence evaluation, global commands to forget will be difficult to utilize. The question of *self-*instruction to forget certain things is an interesting one that has been under-studied, in my opinion. Presumably the global vs. specific F-cue distinction will be at play in the case of self-instruction, though again, most experience is item-istic—i.e., you get evidence that something you recently heard (and believed) is false, so you instruct yourself to forget it; or you have an unpleasant experience that you wish to forget, so you instruct yourself to forget it—in these cases, you will likely fail, and at least implicitly, you will remember the item or experience, and it may play all sorts of unconscious roles in your cognitive system.

Let's turn to some practical lessons we might draw from all of these findings. One practical application is that, when it comes to instructions to disregard information—such as testimony in a trial, for example—it will be more effective to instruct people to forget *everything they have just heard* from a given testimonial source, rather than instruct them to forget specific items or pieces of information. We don't want to introduce a "free rehearsal" of the information we are trying to forget. So what else do we know about how to successfully forget? One small but potentially interesting finding is that of Elmes (1971) that

the word *forget* makes a difference—an instruction to "forget" is more effective than "ignore" or "disregard" or "do not remember", for example. Another, very robust finding (mentioned in passing back in chapter 1, referencing Johnson & Seifert, 1999) is that previously stored information that has been utilized in causal explanations, or in deriving further inferences, is very difficult to root out (i.e., even when the item itself is successfully forgotten, the implications and causal explanations in which it was featured remain).[13] However, if an *alternate* causal explanation is offered to replace the one(s) utilizing the to-be-forgotten information, then forgetting will be much more successful (Golding & Long 1998; Johnson & Seifert, 1999; Seifert, 2002).

> When a compelling account can be offered as an alternative, people are less likely to fall back on misinformation. Replacements are most successful when they account for causal coverage of the events, regardless of their plausibility. When no causal alternative is available, however, it appears to be very difficult to correct misinformation that plays a causal role in an account. (Seifert, 2002: 13)

Additionally, it's easier to forget or replace an item (of misinformation) if not only an alternate causal explanation is offered, but also an explanation for why conflicting reports were presented in the first place. Seifert connects this to the "need for explanation": as "people fall back on explanations they know are wrong in an attempt to "fill in" the causal gap. In a sense, the need for explanation may outweigh the known value of the information" (2002: 13). Explanatory structures interpolate and smooth the information, and thus make it more tractable for processing—they are like hammers that allow the nails to suddenly come into focus.

7.3 The bell that can't be unrung

All of the studies cited above are, ultimately, concerned with *belief perseverance*. This phenomena was one of the motivating factors for questioning standard normative accounts of belief revision—since, in the real world, we apparently not only commonly fail to revise (perseverant) beliefs, *we don't even realize we have them*. Indeed, our belief set may turn out to be glaringly inconsistent, given that we may routinely accept that a previously held belief is wrong, and yet the belief perseveres: we both disbelieve and believe it. Or we explicitly disbelieve it, but act or react *as if* we believe it, suggesting an implicit, and extremely stubborn belief. The studies on *forgetting* show us that it is very difficult to root out anything

that was ever remembered: we very often can't *unring the bell* once it has been rung. And our belief set will include not only some of these un-unringable bells, it will include numerous beliefs that are infected in whole or in part, by the reverberations of these un-unringable bells. These beliefs will be *unrevisable*. We'll be stuck with them, Quinean sensibilities and norms of rationality notwithstanding. This unrevisability of some sorts of beliefs cuts directly against the isotropy that Fodor insists characterizes belief fixation: beliefs are *not* sensitive to all other belief—some are and some are not, depending on the way they were framed during encoding, and reframed during recall. Beliefs *shift* depending on the dimension by which they are being measured.

At the end of PART II of this dissertation, I suggested that what we would look at in this chapter regarding memory should help to *confirm* the account that I laid out in PART II. My claim has been that a belief revision system that could plausibly claim to be tractable and to thereby skirt the frame problem, would need to be predicated on a suite of massively parallel, encapsulated, domain-specific modular structures, including multiple levels of assembled modular integration and interface mechanisms which approximate global, holistic reasoning and deliberation via heuristically-driven processing algorithms. Of course, the only way that such a system could possibly work is to invoke a global processing workspace in which disparate elements can be brought together for processing. And as discussed at length in chapter 6, this global workspace needs to be tractably managed: both in terms of limiting what/how much data is brought in for processing, and (more importantly) in terms of framing and limiting the search procedures engaged to find relevant data in memory. Finally, the manner in which items are stored in memory for later use in processing has to be compartmentalized in such a way as to facilitate those heuristic search procedures: there need to be storage redundancies, and it would be preferable to store compound information in pieces, such that only pieces needed in a given task need be recalled, and pieces which do double-duty (or more) in various compounds can be nested in frames. The result is brute, but effective—"good enough" but prone to be sub-optimal. I laid out 4 ways in which the system would reveal its sub-optimality:

(1) *pervasive unawareness of inconsistency*, due to processing limitations of the amount of data under consideration at a given time;

(2) *cognitive reasoning biases* that would reveal themselves in complex deliberative settings that our systems were not adapted for;

(3) *behavioral, attitudinal, and cognitive effects of false beliefs* that have gone undetected in the system as a result of (1) and (2);

(4) the *generation of pathological beliefs that are incorrigible or irremediable*—beliefs which whose generation is inevitable given functioning of sub-serving modular processes, and which may be (ironically) regenerated and solidified via revision.

Everything that I have looked at in this chapter regarding belief perseverance, false memory, and the near-impossibility of truly "forgetting" what was once remembered, seems to support the conclusions listed above. The upshot of this account is that it completely undermines any hope for a "Quinean" sort of belief revision practice and adherence to standard norms of procedural and epistemic rationality, and it reveals belief to be functionally anisotropic. Our belief revision system has to piggy-back on our systems of memory encoding and retrieval, and those memory systems will inevitably generate some (perhaps many) unrevisable memories, and hence beliefs. And this is by design—not in the sense that unrevisable beliefs are the aim of the system, they are indeed merely a side effect. But the processes and mechanisms that allow for a system that can revise belief *at all* will be limited in ways that make it impossible to revise *all belief*. Consider this the ironic principle of belief revision—*a system capable of revising belief at all must be incapable of revising all belief.* Recall Cherniak's admonition that "global rationality will involve some local irrationality". I would emend this to say: *global revisability will necessitate some local unrevisability*. I want to reiterate once again that this *unrevisability* is *exactly what we should expect* from a modular system—from perceptual illusions on down, we get the same effect: certain contexts will elicit responses that can't be overridden, that are completely, incorrigibly, irremediably impenetrable. Some bells cannot be unrung. Some beliefs cannot be unbelieved.

7.4 Review and look ahead

In this chapter, I turned to empirical findings regarding memory encoding and retrieval to seek confirmation for the account of belief revision I traced in PART II of this dissertation. I have argued here that the evidence of misinformation effects, belief perseverance, and false memory implantation demonstrate exactly the sort of functioning that should be expected from the heuristically-driven and modularly structured cognitive architecture I have defended. Furthermore, the conditions under which *forgetting* is possible lead to the

conclusion that there the cognitive system I have described will *inevitably* result in beliefs that *can't be revised*: they can't be revised largely because they can't be forgotten. These "un-unringable bells"—isolated desert islands of belief—are the systematic pattern of breakdown that a system such as I have described is prone to. I think *all* of us will be subject to some of these irremediable, unrevisable beliefs, and as a result they shouldn't be considered a mark of *irrationality*. If rationality hinges on the eradication of all inconsistencies of belief, then rationality isn't even possible *in principle*, given the human cognitive system

In the final chapter, I turn to delusional beliefs, which are largely considered to epitomize irrational belief. I do so with a twofold aim:

1) I will highlight some syndromes of delusional belief that serve as paradigm examples of unrevisability, even in the face of massive evidential override, and argue that the existence of such belief states will confirm the account of belief revision I have defended. In short, I will argue that delusions are prototypical patterns of systematic breakdown resulting, specifically, from deficits in the modular cross-modal perceptual integration stage;

2) I will utilize my account to help resolve some puzzles and theoretical disputes in the literature on delusion, at least with respect to so-called *monothematic* delusion. I will also suggest a positive research (and potentially treatment) path for some delusional syndromes based on predictions that fall out of applying my account to the subject of delusional belief.

Notes for chapter 7

[1] To repeat the analysis I gave initially in chapter 1: In the study, subjects were given false information about their abilities to perform a particular task (in this case, detecting and distinguishing false suicide notes from legitimate ones)—subjects were told either that they were significantly better or worse than average in this regard. Their judgments about their own abilities at this were formed accordingly. Subsequently, they were debriefed and shown incontrovertible evidence that their "performance" on the task was manipulated, and the information they had been given about their "results" was false, and all beliefs based on it were therefore unwarranted. Subsequent self-reports reflected that subjects *maintained* the belief in their "ability" despite the debriefing.

[2] *Cf.* Thomson & MacLeod (1988) and Devine *et al.* (2000) for comprehensive reviews of evidence from numerous jury studies.

[3] Note that we will return to the prickly issue of "disregard" instructions in the section on intentional forgetting, below, including a lengthy discussion of so-called "ironic process theory" which accounts for exactly this sort of "boomerang effect". See Fein *et al.,* (1997) for a full review of studies on the effects of pre-trial publicity.

[4] Note that the Bugs study resulted in a pretty high hit rate for explicit false recall, but subsequent studies of *implicit* memory were even more impressive, as could be revealed by *implicit attitude* testing which show higher degrees of associations between Bugs and Disney among study participants than among controls. Braun-Latour *et al.* suggest this implies that *semantic* memory for Disney and Bugs has been affected (2004: 19).

[5] Another possibility is that only people prone to *confabulation* will end up with these sorts of rich false memories. This too does not seem to be the case, as a) again, the subjects who "fall for" false memories are not in any other obvious way more prone to confabulation; and b) there is voluminous evidence that all people confabulate to some degree, quite routinely. One of the most documented psychological phenomena is our uncanny ability to lie convincingly to ourselves regarding motivations and intentions to force accordance with subsequent behaviour *a posteriori*, such as the work of Festinger (1957) whose "cognitive dissonance" 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 (Festinger, 1957; cf. Festinger & Carlsmith, 1959). Similarly, there is a wide spectrum of behavior that falls under the heading of *self-deception*. We often act in ways contrary to our professed beliefs and desires, and when called on it, we rationalize and confabulate all sorts of reasons to explain away the apparent contradiction. Sometimes we do it knowingly, but in a large number of cases we are oblivious to it, and sometimes stubbornly refuse to accept we have confabulated (i.e., we accommodate our beliefs to the confabulation). Gazzaniga (1995), working with patients having undergone hemispherectomy—"split brain" patients—also notes that confabulatory behaviour is common when our motivations are opaque to us (note that he cashes it out in terms of a massively modular architecture). Furthermore, Weinstein (1966) and Frith & Nathaniel-James (1996) report that no general memory deficit is found even in cases of schizophrenic patients who routinely engage in (elaborate) confabulation. Hirstein (2005; 2009) suggests that memory deficits are not a necessary part of confabulatory syndromes—even chronic confabulators may perform perfectly well on memory tasks. All this to say: false memory isn't just a problem for *confabulators*—except in the sense that we are all confabulators. False memory is the result of different mechanism. So, it doesn't appear that tendency to adopt false memories as one's one (veridical) memories is particularly connected to memory deficits in general, or to confabulatory proneness in general.

Excess suggestibility is a more plausible candidate. One finding, related by Loftus (2005) is that the very young are more likely to accommodate false memories, as are the very old—both populations that are prone to excess suggestibility. (Of course, that correlation could go both ways: perhaps we only attribute "suggestibility" to those who easily accommodate false memory.) Loftus (2005: 362) reports that her research has revealed a positive correlation between misinformation susceptibility and *empathy*. Emotional valence in general has some effect on whether or not misinformation gets planted. McNally *et al.* (2004), studying people who report experiences of alien abduction, note that the physiological expression of emotion that accompanies the act of "remembering" can be deceptive, in the sense that to the *subject*, the emotional associations to the memory count as a point in favour of its veracity, though there is little actual connection between "the physiological markers of emotion that accompany recollection of a memory [and] evidence of the memory's authenticity" (2004: 496). On the other hand, McGaugh (2004; *cf.* English & Neilson, 2010) showed evidence that increasing the emotional valence of the context *decreased* the chance of memory distortion.

[6] In chapter 2, above. Note that we might quibble with referring to this as a "long-term/short-term" issue. I think "working memory", following Baddeley (1986) would be more precise.

[7] In §7.2.2 below, we will look at Wegner's (1994; 1997) *ironic process theory*—the effect Roediger *et al.* point to here may be an example of that at work. I will return to it in the later section.

[8] See Bjork & Woodward (1973); Bjork & Bjork (2002); Elmes & Wilkinson (1971); Walster *et al.* (1967); Paller (1990); Basden *et al.* (1993); Johnson & Seifert (1999); Wilkes & Leatherbarrow (1988); Golding *et al.* (1990).

[9] Recall Gilbert's claim, discussed above, in chapter 1, that in order to understand something, one must provisionally believe it, and only later can one evaluate it for truth—the *Spinozan* view, as he class it. Perhaps here we see the results of this in terms of memory: mere exposure to certain information gets it into memory in

a way that is subsequently implicitly available—i.e., it is put on the shelf—but only under certain conditions is the memory tagged for explicit retrievability.

[10] Note that this would get us roughly the distinction Gendler is after with *alief* vs. belief, as discussed in chapter 1, earlier.

[11] Gilbert (2011) posits an account somewhat related to Barsalou's *simulation* account of concept formation, employing a dual-systems approach to the generation of predictions—"previews" and "premonitions"—based on combinations of implicit and explicit memories. What he tries to establish are the conditions under which these predictions tend to *fail.* The "previews" we generate *often* go awry because the constraints under which they are generated are not optimal. Four ways in which our previews are not optimal, according to Gilbert (2011: 1337-39):

- ***Previews are unrepresentative***—we generate predictions based on *available* memories, not necessarily typical memories. He cites Kahneman *et al.* (1993)—the *icewater test*—in which subjects whose hands were placed in icewater for an extended time rated the experience as less painful if there was a slight warming towards the end, as opposed to another test in which their hands were in the slightly warmer icewater the whole time.[11] Additional support comes from Morewedge *et al.* (2005) in which subjects are asked variously to "remember a time you missed your train" vs. "remember the WORST time you missed a train"—everyone remembers the *worst.*
- ***Previews are essentialized***—we only remember the essential features of the memory we are basing the preview on. Gilbert cites Schkade & Kahneman (1998) as an example: people who are looking forward to their move to California will be focused on the *sunshine*, having seldom considered the *traffic* in their preview.
- ***Previews are truncated***—we remember beginnings and endings mostly, often neglecting to include middle periods of an event memory. Note that this is a common fact about memory: even in remembering lists of items, or strings of symbols, it is the middle that is often hardest to recall later.
- ***Previews are comparative***—as Gilbert explains, "imaginary chips are compared to imaginary sardines, but real chips are not" (1339). Anchoring and adjustment proceeds quite differently in imaginary contexts than real-world ones.

Gilbert takes it one step further, and in a move reminiscent of Carruthers' (2006a) iterative inference faculty engaged in "cycles of inner speech", Gilbert suggests that System 2 "tries out" the prediction (internally) and checks the System 1 (emotional) reaction. The prediction can then be fine-tuned as needed. Gilbert's analysis highlights the *heuristic*-based nature of mental previews and the memory retrievals they entail. Gilbert cashes this out in dual-systems terms: the conscious previewing and predictive function is necessarily a reflective System 2 affair. But, as we have seen repeatedly, and I have tried to argue above, it would be a computational disaster for System 2 to have *access* to all the nitty gritty stored in memory (and picked up via perception). When System 2 wants to go on a memory search, that search is "prepped" in a sense via System 1 algorithms, which run on heuristics that are fast and frugal. System 2 can quickly generate a feature set from memory, but at the price that this set is based only on available, truncated, essentialized, easily comparable features that will "resonate" in the global workspace (to borrow Jackendoff's terminology—this is not quite what Gilbert would say). System 2 now *uses* that "memory" that is presented to consciousness and can run logical inference principles to draw predictive conclusions. Note that *this*, in itself, is entirely tractable—the inferences themselves are not necessarily a computational challenge at all. As we have seen from the beginning of this dissertation and the introduction of the *frame problem*, the computational challenge is in the search procedures required *prior* to a (normatively acceptable) inference procedure. Once the data upon which to base the prediction is assembled (thanks, System 1 heuristic algorithms!), the inference can run: perhaps it is a slow logical, linear process in consciousness or perhaps it involves even more System 1 implicit work, running a simple Bayesian operation. But it's not courting combinatorial explosion at this point. The context is framed and the relevant information is on hand. At this point all we need is a heuristic halting procedure to judge the process satisfactorily completed.

[12] Another explanation comes from Paller (1990; also Basden *et al.*, 1993) is that items and lists are encoded separately. Neither explains specifically what that entails—just that there is some differentiation in electrophysical correlates in subjects exposed to list vs. item methods of cued-recall. Basden *et al.* merely conclude that this demands differentiated explanatory mechanisms at the encoding stage for lists vs. items.

[13] See also Wilkes (1988); Wilkes & Reynolds (1999); and Seifert (2002) for a full review of these studies and their implications.

8 Delusion

In this final chapter, I turn to the subject of delusional belief with a two-fold aim:

1) I want to focus on monothematic delusions as *support* for the account of modular, heuristically-constrained belief revision that I have given in PART II of this dissertation. I will argue that monothematic delusions are precisely the *predictable and systematic breakdown patterns* one should expect to result from a belief revision process sub-served by multiple levels of integrative modular assemblies.

2) I want to show how my account of belief revision can be *applied* to theoretical disputes regarding delusions, their etiology, doxastic status, and possibility of remediation. I will argue that viewing monothematic delusion as the result of integrative modular "misfires" provides an elegant solution to a number of the puzzles faced in delusion research and treatment.

I will proceed as follows. First, I will very briefly sketch the terrain of current accounts of delusion, in order to highlight some of the primary disagreements as to how to best explain delusion, and how it fits into a broader account of belief. I will examine these disputes mainly in terms of binaries, just to simplify somewhat, and in each case I will state my own position. In §8.2, I am going to look at two fascinating and puzzling monothematic delusional syndromes: the Capgras delusion (in which subjects believe that their loved ones have been replaced by impostors), and the mirrored-self misidentification delusion (in which subjects perceive their own reflection as a stranger). I will give an account that explains the etiology of these delusions as the result of integrative modular functioning, directly analogizing them to the many cross-modal *illusions* that strike the perceptual system, and which we looked at it great detail in PART II of this dissertation. I will also suggest some further experiments and possible treatments for these delusions based on predictions that result from my model. Although I will only provide a detailed discussion of these two delusions in this section, I will suggest that it's quite likely that further, perhaps all, forms of monothematic delusion could be well accounted for using the explanatory power of

integrative and assembled modularity as the primary locus for the delusional belief, and the source of its circumscription and irremediability.

In 8.3, I will look at limitations of my account—namely, that while it may work well for monothematic delusions, it might not be as explanatorily helpful for *elaborated* or *polythematic* delusions, such as persecutory and motivated delusions. I will gesture toward a way to use a modular thesis to account for these, but this will be speculative, and I will not defend it at length. Finally, I will address the issue of the *un-unringable bell*—or in this case, the un-unbelievable belief. I think the account I have given in this dissertation predicts and *requires* that there will be such recalcitrant, irremediable belief states, both in putatively "rational" and floridly "irrational" people. They are a feature of the system, not even necessarily a bug. A system that worked "*better"* wouldn't work at all.

8.1 Theoretical disputes

Current research on delusion is lively and highly interesting, and there are a number of questions as to the nature, doxastic status, etiology, progress and treatability of delusions. A number of recent books have attempted to coordinate the many accounts across disciplines from neuroscience, psychiatry, psychology and philosophy, notably Bortolotti (2010), Coltheart & Davies (2000), Radden (2011), and Bayne & Fernandez (2009). I will follow the lead of Bortolotti in particular in terms of my brief sketch of the theoretical landscape, laying out the disputes largely in binary terms (i.e., issues with roughly "two camps").[1] In this section we will look at 3 distinct questions one might pose regarding delusion, separated in the three subsections to follow.

8.1.1 Doxastic or no?

The first question we might ask about delusions is whether or not they are *beliefs* at all. My own bias on this is likely clear from the introduction to the chapter, as I referred repeatedly to delusional subjects "believing" the contents of their delusion, and it is indeed the case that I want to call delusions *belief* states, specifically *false belief states.*[2] This makes me a "doxasticist" regarding delusion. Doxastic accounts of delusion answer the question posed by this section in the affirmative: delusions *are* belief-states—false ones,

"pathological" ones, "irrational" ones, depending on your theory—but belief states just the same. Non-doxastic accounts, on the other hand, suggest delusions do not meet the definitional criteria of belief, for one reason or another, and should be viewed as something *else*. Gendler (2007) suggests delusions are a sort of "pretense":

> I think the most helpful way of distinguishing beliefs from other related cognitive attitudes is neither through their subjective vivacity, nor through their dispositional connection through desire to action, but through their *telos* of truth, such that the status of beliefs depends upon their being *reality-sensitive* in certain crucial ways… What makes my commitment to P a *belief* that P—as opposed to an imagining or supposition that P—is that my acceptance of P as true is contingent on how I take the world to be: my attitude is one whose fundamental satisfaction conditions require that it have been formed (whether intentionally or not) through the workings of a cognitive system which regulates certain of my cognitions in ways designed to ensure that I bear this attitude only toward truths. (2007: 236)

Currie (2000; *Cf.* Currie & Jureidini, 2001) argues that delusions are closer to *imaginings* than belief states. Currie bases this claim on delusions suffered in the grip of schizophrenia, resulting as a side effect of loss of agency (i.e., feelings of alien control, or voices/thoughts):[3]

> [S]ome symptoms of schizophrenia, such as delusions and hallucinations, involve a loss of the capacity to identify imaginings. This is consequent on a general loss of the sense of agency ... So the explanatory picture looks something like this: a loss of the sense of agency leads directly to such symptoms of schizophrenia as a sense that aliens are in control of one's body. And indirectly, via a loss of the capacity to identify one's imaginings, loss of the sense of agency leads to delusions and hallucinations. (Currie, 2000: 181)

Egan (2008) posits a new category—*bimagination*— suggesting that "delusional subjects don't straightforwardly believe the contents of their delusions, nor do they straightforwardly imagine them. Instead, they bear some intermediate attitude ... with some of the distinctive features of believing, and some of the distinctive features of imagining" (Egan, 2008: 2). Others stress the *experiential* (Gold & Hohwy, 2000), *phenomenological* (Gallagher, 2009) or "bottom-up" (as opposed to "top-down") aspects of delusion. Mundale & Gallagher (2009) describe both views:

> We support a bottom-up model of delusion, one that holds that delusions are immediate and non-inferential. With respect to the noninferential character of delusion, our approach is similar to that espoused by Gold & Hohwy (2000) in which delusions are referred to as 'disorders of experience'. At the same time, however, we acknowledge the explanatory appeal of top-down models of delusion, in which delusions are thought to derive from predictable, cognitive errors… we argue that the kinds of errors to which such top-down models typically appeal may themselves be understood, in certain crucial respects, in a bottom-up way, or as part of the immediate experience. (Mundale & Gallagher, 2009: 513)

All of these alternatives to seeing delusional states as *belief* states are well motivated. To be sure, one of the major reasons for thinking that delusions shouldn't count as beliefs is that delusions simply do not "perform" like beliefs: delusions are very often behaviorally, affectively and cognitively *inert* (Frankish, 2009: 270). A person with the Cotard delusion (Cotard, 1880) sincerely asserts the belief that she *is dead*. Yet, the "belief" doesn't drive *behaviour* in the way belief normally does. In fact, very few behaviours of the person suffering from the Cotard delusion seem to stem from the delusional belief, aside from its repeated assertion. She still gets up, has breakfast, pays for the bus, etc.

Another example: a man experiencing the Capgras delusion asserts that his mother has been replaced by an impostor. But he doesn't go *looking* for his *actual* mother—the one who is missing—instead, he simply sits down with the impostor for Sunday dinner. This seems a strange behavioral oversight. Not only is the belief that mother is an impostor behaviorally inert, it is also apparently *affectively* inert: the man may be *puzzled* by the presence of the impostor, but not particularly *troubled* by it. Another example of affective inertia of delusional belief is in the *mirrored-self misidentification* delusion—a delusion in which the subject perceives his or her own reflection as a *stranger*. In many cases, emotional states often fail to follow appropriately. Breen *et al.* (2001) report the case of subject FE:

> FE believed that his own reflection was another person who was following him around, not only in his home, but anywhere there was a reflecting surface. FE had attempted to communicate with the person on numerous occasions and was somewhat perturbed that the person never replied but was otherwise undisturbed by the stranger's presence… [Additionally], FE's semantic knowledge about mirrors was entirely intact. (Breen *et al.,* 2001: 240)

I think that if it were me, I'd be more than "somewhat perturbed" by this, and not simply because the stranger failed to respond to me.

The last line of the above quotation points to the final way in which delusional beliefs can be dissociated from stereotypical belief: delusional belief is marked by a *cognitive inertia*, insofar as the delusional belief, at least in cases of monothematic delusion, does not interact appropriately with *other beliefs.* Contradictory beliefs are not revised, and the delusional belief does not provoke many (or sometimes any) elaborative inferences or implications—it does not lead to further belief, nor in many cases does it lead to the pruning

of beliefs it is in direct contradiction to. Young & Leafhead (1996) cite the case of a 29 year old woman, JK, suffering the Cotard delusion:

> We asked her, during the period when she claimed to be dead, whether she could feel her heart beat, whether she could feel hot or cold, and whether she could feel her bladder was full. She said she could. We suggested that such feelings surely represented evidence that she was not dead, but alive. JK said that since she had such feelings even though she was dead, they clearly did not represent evidence that she was alive… We then asked JK whether she thought we would be able to feel our hearts beat, to feel hunger, and so on if we were dead. JK said that we wouldn't…JK recognized the logical inconsistency between someone's being dead and yet remaining able to feel and think, but thought that she was none the less in this state. (Young & Leafhead 1996: 158)

She "recognized the logical inconsistency" but clung to her belief just the same. Earlier, I discussed the problem of inconsistency awareness, and how it complicates belief revision. But here we have a case where the subject is perfectly aware of her logically inconsistent beliefs, yet not inclined to resolve the inconsistency. So simple inconsistency unawareness is not the issue here.

All of these are good reasons to suspect that delusional "belief" may be so far removed from what we expect of prototypical belief that it ought not to be included under the same category. However, the intuitive drive to treat delusions as (perhaps defective but nonetheless) beliefs is fairly strong. Buckwalter, Rose, & Turri (2013) have shown fairly convincing evidence that *folk* intuitions treat delusions as prototypical belief, even when the subject to whom the delusion belief is attributed clearly holds contradictory beliefs. Holding contradictory beliefs doesn't mean one of them *isn't a belief*, at least as far as the "folk" are concerned, apparently:

> First, using different measures, we show that the folk readily classify Capgras delusions as *beliefs*. Second, we show that people view these delusions as beliefs because *frequent assertion* is a powerful cue to belief ascription. In folk psychology, frequent assertion *just is* a behavioral pattern stereotypical of belief. In other words, viewed in the ordinary way, there are situations in which frequently asserting Q is true *suffices* for believing Q. Third, delusional patients are readily viewed as holding *contradictory beliefs*, which can explain the ambivalence we feel when considering such cases. (Buckwalter *et al.,* 2013: 6)

Frankish (2009: 271) also comes down on the side of doxasticism, noting that many *non*-delusional beliefs exhibit the same sorts of cognitive, affective, and behavioral inertia at times—especially those beliefs that are "compartmentalized" in the sense that they piggy-back on System 1 processes that are opaque to us, and as a result may enter into contradictory

dyads and triads with other beliefs we have without our being aware. Bayne & Pacherie (2005) also resist the non-doxastic position, arguing that "the cluster of dispositions that mark out particular beliefs typically includes dispositions to certain emotional responses, but we resist the thought that emotional and affective dispositions are constitutive elements of the belief stereotype" (2005: 184). Bortolotti sums up the defense of the doxastic view nicely, noting one of the studies of belief perseverance we looked at earlier—Nisbett & Ross (1980)—as support:

> The most common versions of anti-doxastic arguments seem to rely on an idealization of normal belief states, and impose constraints on delusions that typical beliefs would not meet. The assumption seems to be that beliefs are essentially rational, and that delusions are not beliefs because they are not rational. But the abundant psychological evidence on familiar irrationality tells us that ordinary beliefs are often irrational in exactly the same way as delusions can be – although to a lesser degree. It is sufficient to think about hypocrisy, about prejudiced and superstitious beliefs, and about the many biases that affect belief updating in normal cognition to realize that the same kinds of irrationality that we find in delusions are also common in many ordinary beliefs (e.g., Nisbett and Ross 1980). (Bortolotti, 2009: §4.2)

I think delusions *do* constitute belief states. As I will argue in §8.2, delusional beliefs are generated with and maintained by the same systems that non-pathological beliefs are—and that, even in cases of delusion, those systems are actually operating *as designed*. The problem with delusional belief is a *content* problem, and the content problem shows up as a result of encapsulated modular integrative functions at too low a level to be remediated. This results in beliefs that are defective and atypical in many ways—but beliefs nonetheless.

8.1.2 Explanation or endorsement?

If we go with a doxastic account, as I am inclined to, we have a second question—what is the etiology of the delusional belief? Two sorts of accounts diverge on this question. *Explanationist* accounts treat the delusional belief as an *explanation* of a bizarre or anomalous perceptual experience. Brendan Maher (1974; 1988) typifies the explanationist view:

> Strange events, felt to be significant, demand explanation. It is the core of the present hypothesis that the explanations (i.e., the delusions) of the patient are derived by cognitive activity that is essentially indistinguishable from that employed by non-patients, by scientists, and by people generally. (Maher, 1974: 103)

In this sense, the delusional belief isn't a result of deficient procedural rationality—the cognitive systems that subserve rational thought are operating as they should, according to Maher. One's cognitive system can only process the information it has—and the primary source of information is perceptual experience. In the face of *any* experience, we employ our cognitive systems to *explain* it. The delusional subject simply has an anomalous experience, and *given* that experience, the delusional explanation is not that bizarre. According to Bortolotti (2010), one of the motivations behind the explanationist view is to account for the fact that the content of the delusional belief is more *specific* than the content of the experience that provokes it:

> For instance, in the Capgras delusion, I see a stranger who looks like my father (experience), and I explain the fact that the man looks like my father by coming to believe that he is an impostor (delusion). In persecution, I perceive a man's attitude as hostile (experience), and I explain his looking at me with hostility by coming to believe that he has an intention to harm me (delusion). (Bortolotti, 2010: 31).

Stone & Young (1997) concur, suggesting in their own discussion of the Capgras delusion that "studies suggest these delusions can best be explained in terms of the person suffering from the delusion attempting to make sense of or explain a disturbing perceptual experience that is brought about by brain injury" (1997: 330). Detractors of this sort of view, however, note that in cases like the Capgras delusion, calling the belief that the subject comes up with an *explanation* seems hardly correct, given the bizarre conclusion that is arrived at is so odd as to hardly qualify as 'explanation'. Fine *et al.,* (2005) suggest that the explanationist accounts of the Capgras delusion "explain the anomalous thought in a way that is so far-fetched as to strain the notion of explanation" (2005: 160).[4]

On the other hand, *endorsement* accounts don't posit an "explanation" stage, but rather claim delusions are akin to perceptual beliefs—perceptual experience presents itself *as* [whatever delusional content], and the patient merely *endorses* that belief (consciously accepting, or passively making no or incomplete efforts to override or defeat it).

> According to endorsement models, the experience comprises the very content of the delusion, such that the delusional patient simply believes—that is, doxastically endorses—the content of his or her experiential state or at least something very much like the content of this experiential state. (Pacherie, 2009: 106)

Bayne & Pacherie (2005) highlight what they refer to as "doxastic context-sensitivity", giving the example of Capgras delusion, in which the subject may recognize a spouse over the phone (aural recognition) but not while in view (visually).

> Perhaps the content of the person's visual state is such that it leads them to endorse the impostor hypothesis. When, however, this visual evidence is absent, the person's normal disposition to believe that their spouse is their spouse is triggered. Here, unqualified ascription of any belief concerning the identity of the person's spouse is problematic. Instead, the tempting thing to say is the person's beliefs concerning the identity of their spouse are dependent on their current perceptual information: to a first approximation we might say that the person has the impostor belief when, and only when, he is in visual contact with her. (Bayne & Pacherie, 2005: 185)

What this seems like is that, depending on the sensory modality that is being employed to *determine* recognition, a recognition "belief" will be simply endorsed directly from perception.[5] Note the similarity here to my own discussion of the Müller-Lyer illusion in chapter 3: there I argued that the "judgment" of the lines being "unequal" is a direct endorsement of what perceptual modules are reporting. Even in the case that the illusion has been revealed, and you come to believe the lines actually are equal in length after all, this is only possible after *experiencing* their sameness somehow (masking the distractor arrows, adding a ruler, etc.) and hence you have merely endorsed *that* belief.[6] Delusion, like *illusion*, is a matter of endorsement, not explanation.

8.1.3 One factor or two?

Assuming we can settle the issues of whether delusions count as beliefs (I have said they should) and whether delusional beliefs are an explanation or simple endorsement of an anomalous perceptual experience (I have argued for endorsement), we will still be left with the difficult question as to why, once it has formed, the delusional belief *perseveres*—even in the face of overwhelming evidential override. Delusional beliefs are often considered the gold standard of irrationality precisely because they seem completely immune to revision, impervious to evidence, and impenetrable by other beliefs the delusional subject simultaneously holds true. Whether one takes an explanationist or an endorsement position on delusion, one still needs to answer the question as to why the (often extremely unlikely and bizarre) belief is not immediately rejected after it has been formed.

The prototypical "one-factor" account of delusion is Maher's. Recall that for Maher (1974), the person suffering a delusion is not *irrational*, rather, she is simply doing the best she can to explain and comprehend a defective or anomalous perceptual experience. The experiential deficit is all that is needed to generate the delusion—it is the only factor. On Maher's view, the rest of her cognitive systems are working as they are supposed to, attempting to make sense of this (nonsensical) perceptual experience. Now, I put pressure on the explanationist side of that claim, above, as the "explanation" in cases of bizarre delusions (like Cotard's or Capgras) hardly seems like an appropriate explanation. And the main *defect* of it *qua* explanation is that it won't withstand any scrutiny. Yet delusions *do* withstand scrutiny—indeed, sometimes they withstand *all* levels of rational scrutiny and are impervious to logic and revision.

This sort of pressure on one-factor accounts of delusions leads naturally to the more common move, which is to invoke a *two-factor* model, in which an initial perceptual deficit or anomalous experience (the first factor) is responsible for the genesis of the delusional belief, and a *second* deficit—to the subject's belief formation and evidence evaluation systems—is invoked to explain the fact that the belief is not immediately (or ever) revised or defeated. Langdon & Coltheart (2000) express the view clearly:

> Bizarre delusions, we argue, require at least two deficits: (1) at least one form of perceptual aberration, whether caused by dysfunction of a sensory mechanism or caused by dysfunction of attentional/orienting mechanisms; and (2) a breakdown of normal belief evaluation. Thus, the existence of deficits that cause one or more perceptual aberrations is not sufficient to explain the formation of delusional beliefs. (Langdon & Coltheart, 2000: 213)

The positing of a second factor involved at the level of "normal belief evaluation" is intuitively appealing, and certainly serves to explain unrevisability of delusional beliefs, once formed. Davies *et al.,* (2009) note the syndrome of anosognosia ("denial of illness"), where patients refuse to accept their medical/physical condition (e.g., a recent amputee may suffer the delusion that everything is fine, and the severed limb is still intact). Davies *et al.* note that anosognosics have so much evidence that immediately defeats their delusional belief, that there *must* be a second factor to explain the maintenance of their syndrome.

> In the presence of a first factor, knowledge of paralysis requires a process of discovery that is not especially demanding for cognitively intact individuals. But anosognosia for hemiplegia arises when the first factor is accompanied by additional impairments that impact negatively on observation and inference. (Davies *et al.,* 2009: 197-98)

The second deficit here is chalked up to some impairment to observation and inference.[7] Accounts of the specific inferential impairment vary. According to Garety *et al.* (1991) the problem is that delusional subjects are especially prone to *jump to conclusions* (the *JTC* bias). McKay *et al.* (2007; 2005) offer the suggestion that motivational biases may leave delusional beliefs improperly evaluated, as "individuals prone to the *second* factor are misled when forming beliefs, such that beliefs formed are increasingly congruent with wishes and increasingly incongruent with reality" (McKay *et al.,* 2005: 323). Langdon & Coltheart (2000) argue that the second factor is an attributional bias to "favour personal-level causal explanations over subpersonal-level causal explanations" (2000: 196). Stone & Young (1997) argue similarly that "there is the challenge of balancing observational adequacy with conservatism. In... delusion, the balance goes too far in the direction of observational adequacy. The important point is that this is a matter of balance (hence of bias), not of deficit" (1997: 350). Bentall *et al.* (1994) propose a model for persecutory delusions in which motivational biases constitute a second factor, as "in deluded patients, explicit activation of self-ideal discrepancies by threat-related information triggers defensive explanatory biases, which have the function of reducing the self-ideal discrepancies but result in persecutory ideation" (1994: 339) Davies, Coltheart, Langdon & Breen (2001) also offer a further argument in favour of the second factor based on the fact that not *everyone* with the underlying (first factor) deficit develops the delusion. They note that

> On Maher's view, simply suffering from any one of these experiences would be sufficient to produce a delusion, because a delusion is the normal response to such unusual experiences. It follows that anyone who has suffered neuropsychological damage that reduces the affective response to faces should exhibit the Capgras delusion; anyone with a right hemisphere lesion that paralyzes the left limbs and leaves the subject with a sense that the limbs are alien should deny ownership of the limbs; anyone with a loss of the ability to interact fluently with mirrors should exhibit mirrored-self misidentification, and so on. However, these predictions from Maher's theory are clearly falsified by examples from the neuropsychological literature. (Davies *et al.,* 2001: 144)

One straightforward reason to doubt the two-factor model, however, is that it generates a number of predictions that do not seem supported at all by case studies of delusion. If the delusional subject has a defective system of belief formation or evaluation, or attends to evidence in a biased or epistemically irrational matter, then we should expect this deficit to show up *all the time.* But this is not the case, especially when it comes to

monothematic delusions, which are severely circumscribed—if there is a second factor deficit in reasoning or belief revision, why does it seem to only show up in the *one, highly restricted, domain* of the delusion? The second deficit, if there is one, should generalize. But it doesn't appear to generalize, which casts doubt on the second deficit. Indeed, the delusional person *can* in fact, often, engage in perfectly acceptable reasoning *about her delusion* in the clinician's office. This is the phenomenon of clinical *insight* (Sackheim, 2004; *cf.* Amador & David, 2004; Bota *et al.,* 2006; Kiersky, 2004). *Poor* insight is common in many psychological disorders: it is the "seeming indifference or unawareness many of the patients display in regard to their own illness" (Amador & Kronengold, 2004). However, even patients in the grip of delusional syndromes can have periods of insight, in which they understand that their beliefs *are* "irrational", and that the evidence does not accord with the belief, and that if the patient's and doctor's roles were reversed, the patient would think the (delusional) doctor to be irrational and the doctor's beliefs to be false. This doesn't seem to fit with the two-factor account, unless that second factor comes and goes in mysterious ways. I will argue below for a one-factor account, at least for monothematic delusions. Whatever "deficits" to belief formation and evaluation systems that may play a *secondary* role in maintenance of the delusional belief, I will argue, are not *deficits.* Rather, they are simply the sub-optimal results of a non-defective system working as designed. The delusional person's belief evaluation system works in basically the same way that it does in the non-delusional person. Given the model of belief revision I have defended in the latter half of this dissertation, *everyone* has a belief formation and revision system that will orphan some beliefs as perseverant and unrevisable—as bells that cannot be unrung. We will be able to explain away the need for a second factor to explain delusional belief perseverance using my account of belief revision in standard cases.

8.2 A tale of two delusions

In this section, I am going to apply the account of belief revision I have defended in this dissertation to the question of monothematic delusion. My mini-thesis here is very simple, and I think somewhat unique: monothematic delusions are the inevitable result of encapsulated perceptual integrative modules trying to "make sense" of corrupted cross-modal

perceptual input. I place "make sense" in scare quotes, as I do not intend this to involve any cogitation—my account is not explanationist in the sense discussed above.[8] I am going to argue that monothematic delusions—at least the two I will discuss directly in this section: the Capgras delusion and the mirrored-self misidentification delusion—are integrative module misfires, directly analogous to the McGurk Effect that was discussed at length in chapter 3 and 4 above.

8.2.1 Modularity and monothematic delusion

I am certainly not the first to attempt to apply the modularity thesis to help explain the etiology or content of delusions—what I am going to argue is original in the account below is simply that it is a cross-modal *integration* stage module that is the culprit. Elizabeth Pacherie (2009) presents a modular account of the Capgras delusion (as I will below as well). She highlights the fact that "[i]f we take as our guide the set of criteria proposed by Fodor (1983) for modularity, it seems pretty obvious that the processes through which feelings of facial familiarity are generated qualify as modular" (Pacherie, 2009: 114). Where I think Pacherie's account doesn't go *far enough*, is in limiting the modularity in this case to the affective domain of familiarity. I will locate the deficit that leads to the anomalous experience (and thence to the impostor belief) in the modular integration phase of face and familiarity perception. Furthermore, I will explain the maintenance of the delusion by invoking modular belief revision practices (according to the account I have defended throughout the latter half of this dissertation).[9]

I believe the key to tying the modularity thesis to an account of delusion is in the invocation of modular assemblies that are *virtually* encapsulated, in the sense that they comprise a closed loop of processing, inheriting all the computational limitations, circumscribed accessibility relations, and domain-specificity that this entails. In particular, I think we can plausibly lay at least *monothematic* delusion squarely at the door of second-level cross-modal perceptual integration modules—the level of integration I discussed at length in chapter 3, above. Recall the examples of so-called "cognitive penetration" of perceptual modules that were discussed in that chapter, which I explained as being not penetrations at all, but rather proof of encapsulated integration of perceptual scenes. The "cognitive penetrations" are, rather, artifacts—misfires—of those integrative systems. I will

argue that monothematic delusions are also. There is a direct, and I think conclusive, analogy to be made.

A further point in favour of a modular account of delusion is that I believe it is the *only* way, in the case of monothematic delusion, to explain how severely circumscribed the delusional belief often ends up. I already discussed above, citing Frankish, the three ways in which delusional beliefs are often *inert*: affecting neither other cognitive states, nor affective states, nor behaviour in the way belief normally would. For example, if I *truly believed* my mother had been replaced by an impostor, I should go looking for my real mother, who is apparently missing; and if I *truly believed* that there was a stranger in my bathroom mirror, I should be pretty scared, and not simply shrug it off as a bit weird; etc. I think the modular thesis applied to delusion explains the circumscription extremely well: the delusional belief exists *only* at the level of explicit assertion (at least in monothematic cases). All other behavioral and affective systems have no access to that belief, as it emerges from a (defective) closed loop of processing, and ends up in a cognitive cul de sac (as I will explain in more detail below). Not even other cognitive systems can make contact with it: it is both a product of and a hostage to the informational encapsulation and domain specificity of the system that generates it. A monothematic delusion belief *gains no traction*—no "resonance" in Jackendoff's terminology—within the cognitive system. The only system that ever picks it up, in many cases, is the language production system, to assert it. And *even then*, it's such a useless (wrong) belief that "telling it to yourself" *still* fails to activate other systems—it's a cognitive orphan. An account of such beliefs predicated on their generation through integrative modular processing structures that handle corrupted (or incomplete) input from lower levels can explain the *orphaned* status of monothematic delusional belief, and the circumscribed inertia that it demonstrates.

Finally, the modular account explains the failure to revise delusional beliefs—using exactly the same explanatory structure I have defended throughout this dissertation to account for and explain belief perseverance in general. Delusional beliefs persevere like any other (mis)belief might persevere. And the particulars of the *way* in which I will argue that delusions are generated in the modular architecture—specifically at the level of cross-modal perceptual integration modules which are fed (mis)matched stimuli from lower level perceptual modules—makes it such that all attempts to review the delusional belief will be

more likely to result in reaffirmation than disabuse. More on the specifics of how the modular thesis applies in the next two sections, beginning with the Capgras delusion.

8.2.2 The Capgras delusion

The Capgras delusion is really quite bizarre. The subject of the delusion sincerely asserts the belief that familiar loved ones have been replaced by impostors.[10] The standard account of the Capgras delusion has it that there is some sort of anomalous perception of *looks like [familiar person X] but it's not [familiar person X]*. This anomalous perception demands "explanation", hence the *impostor* hypothesis is formed, and (somehow) subsequently rises to the level of belief, despite sheer implausibility and massive evidential override.[11] The Capgras delusion involves one physical deficit that is empirically documented, and fairly well-accepted as being a crucial aspect of the etiology of the resulting belief: a deficit in the *covert recognition* system that registers *familiarity.* Ellis & Lewis (2001) have demonstrated a clear double dissociation between the Capgras delusion and prosopagnosia (generally known as *faceblindness*). Recall the discussion in §3.1.3, above, of prosopagnosia: there, I cited Frith (2007) on the potential modularity of the face recognition system, given its hyper-fast automatic processing and ability to identify faces even under extremely noisy or degraded circumstances. Ellis & Lewis (2001; also Ellis & Young, 1990; Breen *et al.,* 2000; Ellis, 1996; Ellis *et al.,* 1993, 1997) have shown evidence that this facial recognition system actually involves the coordination of two systems: one for *overt* and one for *covert* recognition. The overt recognition path relies on consciously perceived sensory information and the cross-referencing of those inputs with memories of the person in question. The covert path takes place in the autonomic system, beneath the level of conscious access, but shows up in increased skin conductance response levels in the presence of familiars—what Ellis & Young refer to as the "glow" of familiarity. In faceblind subjects, areas of the brain associated with overt recognition are not working properly, but the covert system seems to be fine, as evidenced by the "glow of familiarity" that accompanies the (overtly unidentified) person. As a result of this, Ellis & Young (1990) predicted that sufferers of the Capgras delusion would show exactly the opposite results on the same tests: namely, their overt facial recognition system would be operating normally, showing no signs of damage, but the *covert* system would be essentially offline. This prediction was borne out

in numerous studies showing precisely that: Capgras sufferers do not register SCR levels commensurate with familiarity. And this fits perfectly with the reported phenomenological experience of Capgras: the subject asserts that the woman in front of him *looks* like mother, but *isn't* mother. The overt recognition system outputs "looks like mother", the covert system essentially reports "doesn't *feel* like someone we know", yet mother *should* feel familiar, and hence the conclusion is "this must be an impostor". Ellis & Young conclude

> that the affective response system and the personal information must each feed into an integrative device. Such a device would then compare the expected affective response with the actual affective response and some kind of attribution process would take place. How such an integrative device would compare the two forms of information and the workings of the attributional process remains to be understood... (Ellis & Young, 2001: 154)

A "integrative device" is exactly what I will posit in the argument below, however I think the model I will give can give some insight into the "how" that Ellis & Young suggest is unanswered in their own model.

Recall briefly what my argument was regarding the McGurk Effect. I suggested that what is going on in the McGurk illusion *cannot* be construed as the visual system correcting (or penetrating) the auditory system, and the proof of that is in both the *wrongness* of the "correction" and the uniformity of the wrongness across the population. It is totally implausible that everyone "gets it wrong" in the same way unless that wrong answer is the result of an encapsulated process. I argued that a visually mediated lip-reading module and an aurally-mediated phonological parsing module are integrated in a fusing module that cross-checks the signals, interpolates gaps, and corrects for noise, etc. by matching the inputs from both systems. In the real world, when a face is talking to you, the visual and auditory signals will *actually be matched*, so the integration can easily smooth over gaps in both signals to prepare a unified cross-modal representation of the distal stimulus, and pass that on for further processing (presumably off to a semantic interface). Now, since the McGurk stimulus is entirely *unnatural*, the integration system doesn't "know" how to deal with the mismatch, so it does its preprogrammed best: I suggested it likely prioritizes the visual cue, at least as far as letting that rule out what the sound *can't* be (a bilabial) and from there automatically assigns the "closest" match. Notably, you *can* hear the sound correctly by averting your eyes, because in that case the auditory signal goes to the interface stage with

nothing to interface with—so it gets a "free pass" on to the next level: there is no "correction".

I think that the Capgras delusion is *directly analogous* to a McGurk style cross-modal illusion—the only *difference* with delusion is the source of the mismatched cue. The mismatch in the McGurk case is artificial—it happens with a trick video in which the wrong sound has been matched to the lips under view. In the case of the Capgras delusion, the "mismatch" in question is the result of a deficit to one of the lower level perceptual modules that feed that integrator (in this case, the covert recognition system—the *familiarity detector)*. Here's how I think it goes: a man suffering from Capgras is presented with his mother, who he claims is an impostor. His overt face recognition system is working just fine, and he is capable of recognizing that this woman in front of him *looks* just like mother. But his covert recognition system is malfunctioning, and not registering the feeling of familiarity (confirmed via SCR tests). The standard account, at this point, suggests that our patient *generates* the belief "she's an impostor" as an explanation of the mismatched recognition perceptions. But my question regarding this is directly analogous to the question I asked in the McGurk case about the uniformity of the mismatch "correction": in the Capgras case, why do *all the patients with the mismatch arrive at the same conclusion?*[12] The uniformity of the response suggests, to me, no alternative but to conclude that it is the *only* response the system can find to bridge the gap. It is not a hypothesis generated as an explanation—it is an automated response that is immediately endorsed. "Impostor" is to mismatched overt and covert recognition as "/da/" is to mismatched visual /ga/ and auditory /ba/. The "impostor" output from the recognition system is its best guess, and it is its *only* guess. And given that it is an encapsulated modular assembly, it will make that same best guess *every time* it is fed the mismatch, just like in the McGurk illusion.[13]

Below is a diagram to represent the integration "module" in this case, as I am describing it. You can look back to §3.3 to compare this diagram to my McGurk diagram to see the similarities.

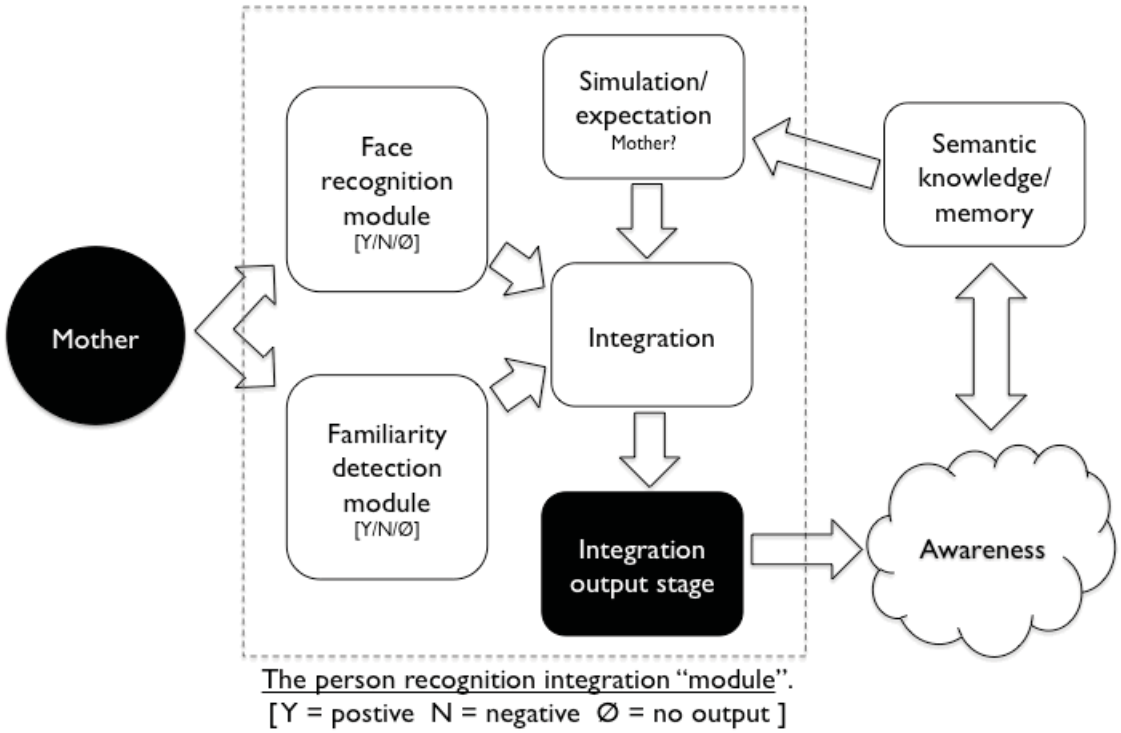


Fig. 10: The person recognition integration "module".

What you should see that's added here is a feedback loop from awareness through semantic knowledge/memory to some sort of simulation mechanism (as discussed above variously by Barsalou and Gilbert). The reason for adding that is that I need to be able to explain the truly troubling part of the Capgras delusion: the fact that the belief *sticks.* The story of the generation of the belief is not that controversial sounding: it's actually not totally crazy that we would have a system that interprets an overt recognition crossed with covert non-recognition as "impostor".[14] The hard part to explain is how that belief survives all the contradictory evidence aligned against it. Even on an endorsement account of the delusion (like Pacherie, 2009) where the belief comes to consciousness with the rich content in place (which is what I am saying is the case: it's not an explanation, the content is generated subdoxastically), one still has to explain why the endorsement isn't immediately rescinded in the face of massive evidential defeat. You might call me out on my analogy to the McGurk Effect—in that case, you *don't* cling stubbornly to the belief that you are hearing /da/ after it

has been revealed as a manipulated video and an illusion. With McGurk, knowing the nature of the illusion, I can experience it, "hear" the /da/, and yet *know* and *believe* that it's false. Why not the same in the Capgras case, if my argument that they are directly analogous holds? Why can't the therapist *explain* the deficit—the brain damage—so that the Capgras sufferer, despite *feeling* that mother is an imposter, can still *know* and *believe* that she's not?

The standard response here would be to invoke the second factor: some sort of deficit in evaluative practice or epistemic rationality. Young (2008) argues that, although "neurological damage could be responsible for an *initial* feeling of unease towards the significant other", that initial feeling will only "transform into a full-blown 'impostor' experience as and when the patient engages in further, dysfunctional cognitive processing" (Young, 2008: 179). Davies *et al.,* (2001) conclude that the unusual perceptual experience of the Capgras subject *alone* cannot account for the full-blown acceptance of the belief, they conclude that there must be an additional deficit in belief formation mechanisms. Ellis & Young (2001) similarly suggest that "it would appear that simply a lack of autonomic response is not itself sufficient to produce the Capgras delusion" (2001: 155). Stone & Young (1997) suggest that the second factor is simply a bias towards "observational" data over considerations of theoretical conservatism:

> The deficit to the perceptual system of those who suffer from the Capgras delusion leads to an anomalous perceptual experience. In the face of this experience, there is the challenge of balancing observational adequacy with conservatism. In people who resolve this by forming the Capgras delusion, the balance goes too far in the direction of observational adequacy. The important point is that this is a matter of balance (hence of bias), not of deficit. (Stone & Young, 1997: 350)

These are all well-reasoned proposals. However, I have already declared myself a one-factor sympathizer in the previous section. I think the two factor account won't work, especially in the case of monothematic delusion, because of the fact that the second factor (reasoning deficit) apparently doesn't generalize *at all*: the defective belief formation practice turns out to be hyper-specifically defective with regard to *only* the one belief in question. It just seems utterly implausible that a person could fail *so massively* as to accept the truly bizarre proposition that mother is an impostor, despite all evidence to the contrary, including their own non-congruent affective and behavioral states, and yet have that massive failure *not* generalize to anything else. It's possible the person subject to the Capgras delusion has no other bizarre beliefs at all. So the second factor, if there is one, is completely localized to

this one delusion. Which means either there is a separate, encapsulated sub-module in charge of the evaluation and testing of *only impostor hypotheses and nothing else*, or this is not a second factor issue. I think it's the latter. Davies & Davies (2009) seem to support the same non-explanationist conclusion for Capgras:

> We now assume that the representational content of the Capgras patient's experience is more specific than 'This is someone who looks just like my wife *but there is something odd about her'* It is, rather, 'This is someone who looks just like my wife *but it is not really her'* (Davies & Davies, 2009: 302).

So we're back to the puzzle: if there is no second factor, why does the Capgras belief stick, whereas the McGurk "belief" doesn't? One reason is that the Capgras belief is the product of a mismatch that *cannot* be disassembled for experience. With the McGurk illusion, the "proof" that *you aren't actually hearing /da/* is to replay the video and *not look.* When you do so, the actual auditory signal is experientially revealed, as the integrative mismatch is blocked, and the automated "misfire" never happens. Sometimes, to be convinced, people have to look back and forth from the McGurk video multiple times to prove it to themselves. But it works, and the *reason* it works to "prove" that the perceptual experience of /da/ is "wrong" is because *it can be experientially replaced*. Recall the discussion of *belief perseverance* and the *misinformation effect* from chapter 7: there we learned that the most effective way to defeat a false perseverant belief—to intentionally forget it—is to *replace* it with one that provided an alternate explanatory structure and took its place in any causal arrangements or elaborative inferences that had been made with or from the faulty belief. I think it's the same in the case of the *perceptual* belief one generates from the McGurk illusion: it needs to be replaced *perceptually* to be replaced at all. Imagine experiencing the McGurk illusion, and when told it's illusion—that what you are hearing is actually /ba/ not /da/—you were *not* able to try out the video again without looking. What if you were just *told* "it was an illusion" and it was explained to you "we overdubbed a video of a guy saying /ga/ with the sound of him saying /ba/; there was no /da/ there." Would you believe it? Or would you insist that you heard /da/ *until you experienced the /ba/*?[15]

With the Capgras delusion, we have no way to experientially disassemble the "illusion". We cannot simply reintroduce Mother with the covert system back online—it's broken. We *can* have Mother call on the phone, or show a video of her, and she will be recognized *as Mother*. But the moment she is in the patient's physical presence, if his eyes

are open, the integrator will output IMPOSTOR. You can't *replace the experience.* So he believes it. Worse, you keep telling him "it's not true, it *can't be true*, it makes no sense"—and in his moments of clinical insight, he agrees. It makes no sense. He knows that, because he *is* rational. But then you present him with Mother, and his recognition system says NOT MOTHER. Every repetition of the experience revalidates the belief that was formed the first time. Not only does it not get overridden or replaced, it gets solidified.[16] In fact, your insistence on telling him the opposite of what he experiences is more likely to lead to his eventually deciding that *you* are lying, rather than that his *experience* is. It doesn't matter that there is a feedback loop through semantic background knowledge—that he gets a simulation that this *should* be mother. When fused with the perceptual inputs, what *should be* gets defeated by what the system is declaring *is*. Semantic knowledge can be employed to help contextualize perception, but it won't override it. This is the lesson we learn from *every optical illusion.* So I have drawn the background knowledge loop into my diagram above to illustrate this: all that background knowledge can do is instigate a re-querying the system ("Isn't that mother? It *should* be... Maybe look again..."), and the system checks and comes up with IMPOSTOR. Pacherie (2008) also argues along these lines, citing Hohwy & Rosenberg (2005) on this point:

> As Hohwy and Rosenberg argue, when the experience occurs in sensory modalities or at processing stages that keep giving the same results and when further intermodal testing cannot be performed (or, if performed, cannot outweigh the results of the dominant modality), it will be taken as veridical. if the experience is generated in a modular way and the module is damaged, this first checking procedure is useless or, rather, instead of helping falsify the experience-based belief, *it will bring only further confirmation of it.* (Pacherie, 2008: 118, emphasis mine.)

Pacherie concludes that "the Capgras patient is not epistemically incompetent; rather, in a way, he is the victim of a vicious epistemic circle" (2008: 120).[17] This is exactly the same conclusion I am arguing for: a non-explanationist sort of Maherianism.

So, even if this integrative modular model is accepted as a plausible account of the etiology and maintenance of the Capgras delusion, there is still the question of why the belief is so bizarrely circumscribed. As noted above, the man who thinks his mother is an impostor nevertheless fails to look for his *real* mother (who is missing!) and does not seem particularly *upset* about what should be quite distressing.[18] This, I would argue, is even further proof that the problem is modular in origin: it's very possible that the encapsulated integration stage I

posit is part of a circuit that sub-serves conscious awareness exclusively. The subconscious, subdoxastic systems that take care of various reactive, behavioral, affective, autonomic functions, may have no need of the complex fused overt-covert recognition representation—they may get all their input directly from lower-level source modules, specific to their domain of function. If my suggestion above that the covert system is the more evolutionarily ancient recognition system, then it makes perfect sense that autonomic, affective, reactive systems would take their cue from *it* rather than from *overt* recognition.[19]

Think of what's going on in the Capgras patient: an integrative system spits out this "impostor" response, which is picked up by awareness (which has no way to override it, and hence has to simply endorse it whenever *in the presence of the stimulus that generates it*). The patient *isn't* "reacting" appropriately, however, because all of his other subdoxastic systems are getting *no "impostor" message*. Those systems don't take their cues directly from conscious awareness. Conscious awareness arguably directs them by marshaling attentional resources to trigger processing that in turn activates various systems to restart the very process that led to the belief in the first place. But the "impostor" belief has no effect anywhere. It makes it to consciousness—it gets thrown onto the blackboard of the global workspace—but it resonates with nothing. The only system that picks it up is language production to *assert it*. But other than that, it's lost in a cul de sac. In the end, it's a belief that literally serves no purpose. Just like hearing /da/ in the McGurk illusion serves no purpose. It's an artifact of a system that *does* serve many valuable purposes, when it works, but in this case, it's misfiring—in a systematic fashion.

8.2.3 The mirrored-self misidentification delusion

Now I want to turn to a second monothematic delusion, giving it the same sort of treatment: the mirrored-self misidentification delusion (MSM), in which sufferers are unable to recognize their own reflection in a mirror, and come to believe it is actually a stranger. It's a fascinating delusion specifically because of how extremely circumscribed it is. The subject of the MSM delusion *does* understand how mirrors work—it is not a simple case of *mirror agnosia* (Breen *et al.,* 2001). Additionally, it's not simply a *facial recognition* deficit—the subject of the delusion often has perfectly normal facial recognition abilities, including for photographs of her own face.[20] In the literature on the MSM delusion, there is actually very

little consensus, and sometimes not even much speculation, regarding how the belief is generated. There is not a well-established physical deficit that can be causally linked to the delusion, as in Capgras. In fact, the MSM delusion can be induced in healthy people via hypnotic suggestion (Connors *et al.,*2012; Barnier *et al.,* 2010), which suggests that no underlying physical deficit is necessary.

I would argue this is another monothematic delusion that is likely amenable to an etiological story based on encapsulated integrative modularity. My main reasons for thinking this are, *again,* the circumscription, isolation, indefeasibility, and most importantly the *uniformity* of the bizarre belief that results: *there is a stranger in the mirror.* I think the crucial studies on the MSM delusion are those that induced the delusion in "healthy" subjects via hypnotic suggestion (Connors *et al.*, 2012; Barnier *et al.,* 2011). In the Connors *et al.* study, subjects under hypnosis were given the suggestion that "you will see a face you do not recognize"—from this subjects generate the "stranger" hypothesis and uncritically endorse it. In this study, healthy non-delusional participants, all of whom we presume do *not* have any lesion in their perceptual system or understanding of self or mirror, converge on the *same* bizarre hypothesis to "explain" their anomalous experience?[21] It's not plausible that it's a cognitive explanation—so it must be an endorsement of a belief that comes *with content*. And again, we need to explain, *why does everyone end up with the same content?* The uniformity of the content suggests a uniformity of output from a dedicated system—just like in the Capgras case, and just like in the McGurk case. Again, all things considered, a system that incorrectly registers "stranger in the mirror" when it gets confused perceptual information is a *safer bet* than one that incorrectly registers "me in the mirror".[22]

I have a proposal for the actual mechanisms underlying the integrative "module" in the case of the MSM delusion. Admittedly, this is largely speculative—though all of the aspects I describe in this case are, I think, well-supported enough in the literature to be plausible. What I will describe is a sort of virtual *mirrored-self recognition module (MSRM)*. I say "virtual" in the sense that this is not a single dedicated processor, but rather an assembly of subcomponent modules whose interaction in this case constitutes an encapsulated, domain-specific, impenetrable loop of processing.[23] When it's functioning correctly, the *MSRM* integrates the outputs of various subcomponent modules. I propose it's a function of a comparator system that fuses input from *intentionality detection* (a system defended by

Baron-Cohen, in chapter 4 above) and *facial recognition,* and then compares that information with expectations based on outputs from motor control. It would go basically like this: the intentionality detection system, faced with one's reflection *doesn't know it's a reflection*—the *ID* system has no "knowledge" of mirrors, nor of faces, let alone how the two may or may not relate—so it simply reports INTENTIONAL CREATURE AHEAD. Meanwhile, facial recognition is working normally and reports LOOKS LIKE YOU. Finally, and crucially, one's *motor control* system produces an *efference copy* of motions one is making, so that other systems can take the *corollary discharge* into account.

A quick explanation of the *efference copy/corollary discharge* mechanism is probably in order before continuing. A well-supported idea in cognitive neuroscience is that motor control systems essentially "copy" perceptual systems on motor commands, so that perceptual systems can take that information into account such that we "can readily distinguish between sensations that are produced by our own movements and sensations that are caused by a change in the environment" (Blakemore *et al.,* 2000: R11). A great example of this is the fact that you generally *can't tickle yourself*, no matter how ticklish you are. This is because the message from your own motor systems inhibits the response. Another, highly important, function of this system, is to correct visual information to account for subjective movement. Presumably when you walk down the street, your head, and along with it your eyes, bobs up and down a bit, depending on how elaborately you strut. As a result, the visual perceptual scene you receive is bobbing up and down. Yet you perceive it as static because your motor control systems have "reported" your subjective motion, and the perceptual scene is corrected in that context.[24] An efferent copy is sent to the simulator to calculate the *corollary discharge*—i.e., the sensory consequences that would/should result from that motion, so that they can be accounted for.

Back to the MSM delusion: I am proposing that when the outputs of these three subsystems—intentionality detection, facial recognition, and the simulation/prediction of corollary discharge based on the efferent copy from motor control—come together for integration, in a normal way, then normal mirrored-self recognition takes place (automatically). In the normal case, the ID system says INTENTIONAL CREATURE, face recognition says LOOKS LIKE ME and motor control provides context to separate out one's own movements from movement of the environment, and crucially, *inhibit* the

intentionality detector from alerting other systems that there is an *intentional other* present. When these are fused, the automated, encapsulated integrator spits out ME IN THE MIRROR. We can add a semantic knowledge base (including knowledge about mirrors) into the loop if we want to as an added layer of cross-reference for the assembled system. Refer to the figure below:

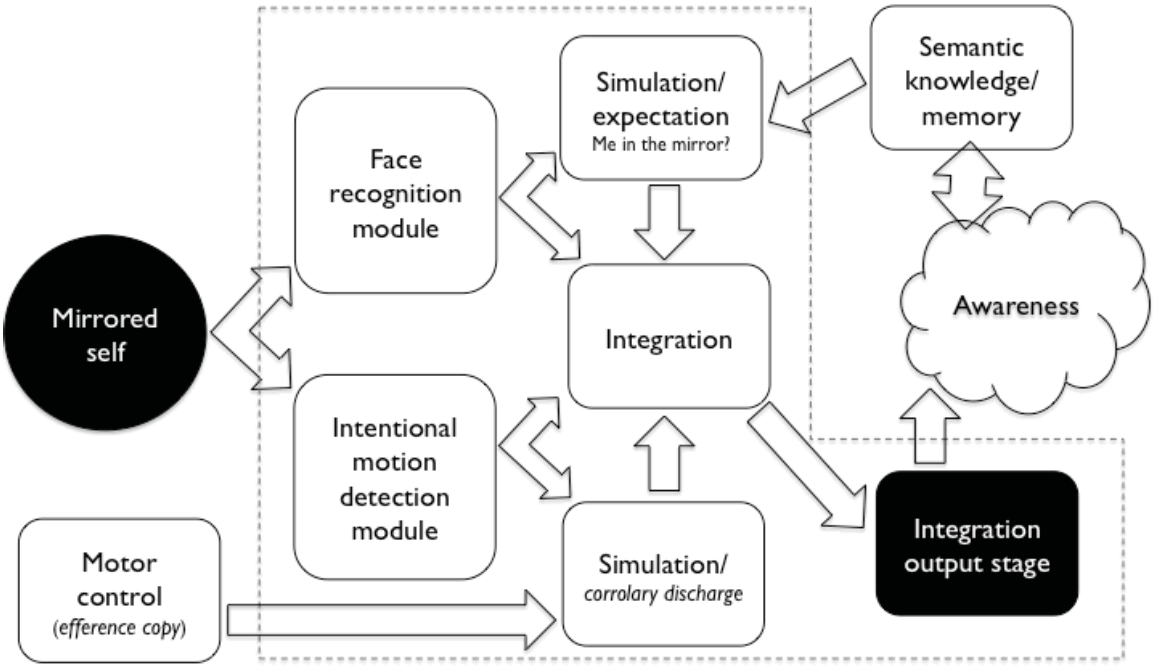


Fig. 11: The mirrored self recognition "module".

My story of the misidentification delusion would be that there is some sort of deficit in mechanism that takes the efferent copy from motor control into account: the fusing of that information with the rest is disrupted, in a way that does not allow the INTENTIONAL CREATURE AHEAD to be merged with ME. In this case, it's plausible that the system opts for a failsafe response, and goes with INTENTIONAL CREATURE AHEAD, dropping ME, and reports out accordingly. No amount of logic or knowledge will intrude on that process as the efferent copy from motor control is not being fused properly, yet that is *the only way* visible movements from another creature can be perceived *as* coming from oneself. Without that, the default is set to OTHER.

As far as I know, no one has subjected patients with the MSM delusion to *further*

tests to see if their ability to "correct" scenes based on motor control operations is compromised or defective in any way. One simple test that could be revealing, is to see if a person in the grip of the MSM delusion *can tickle himself* or cognitively correct for motion-induced fluctuations to visual stimuli. If the answer is yes, then I would suggest that the delusion is a side-effect of a motor control system that is not properly issuing efferent copies of motor commands.

Another experiment that I don't believe has been done, but which could be revealing, would be to hook up a camera and a screen such that the subject is looking at himself onscreen, as if in a mirror. If the feed is updating in real time (i.e., no delay), then this should perform *exactly* as a mirror, and the delusion should be triggered. Now, modulate the video signal so that it is increasingly delayed (i.e., it lags behind), and determine if there is a point at which the mirrored-self becomes recognizable. Recall, that people with MSM *can*, generally identify pictures of themselves, just not *real-time mirrorings*.[25] Perhaps the proposed experiment could reveal a crucial time-delay interval in which the mirrored-self image and the efference copy-generated expectation need to overlap. And perhaps in the deluded subject, either the overlapping in cognition is delayed, or, alternatively, the interval is set too narrowly, such that the speed of processing the integration *is not fast enough* to "make it" before the overlay window is closed. Another possibility is that the integrator *itself* is defective, and can't properly fuse the information, so it defaults to a failsafe: *stranger*. Recall the hypnotically-induced cases of MSM: we assumed these subjects had no *physical* deficit in the system, yet the delusion was induced from the suggestion that they "would not recognize the person they saw". So perhaps simply blocking facial recognition from getting to the integrator confuses the system so that it reverts to the failsafe output.

Regardless of the specific locus of the underlying deficit, I think the MSM delusion is the inevitable result of a breakdown in a modular, integrative assembly, just as in the case of the Capgras delusion, and in the case of cross-modal sensory illusions. It can't be corrected unless corrected experientially (which is impossible as long as the deficit is operative). Perhaps my video feed time-delay experiment could be used therapeutically, actually—one could dial up and down the delay on the video feed while doing active therapy and talking through the experience as it is modulated—but that's totally speculative. The point is that the

delusion has all the hallmarks of a modular breakdown: the pattern is systematic (i.e., people with the anomalous perceptual experience all converge on the *same* bizarre false belief); the belief is perseverant despite contradictory evidence (but importantly, not *experiential* evidence, which is what is needed to replace the belief); the belief is severely isolated, and not being picked up by any other system (except language for assertion) implying structural or representational localization; and the system that sub-serves mirrored-self recognition (normally) has a clear ontogenetic path.[26]

Do we need a "second deficit" to explain why the "stranger" hypothesis is not subsequently overridden? I do not think so, for the same reasons as in the Capgras case: there is no *deficit* in the belief revision or rational evaluation system that licenses the perseverance of the deluded MSM belief. The system, under normal conditions, works to preserve—and fails to override—*any* belief that encapsulated perceptual subsystems serve up, *unless* it can be *replaced* with the right sort of content. Just like with an optical illusion: you can "know" it's an illusion, but you would never *believe* that just because someone told you, or proved it logically. The only way to come to know that a perceptual illusion is an illusion is to have it revealed to you *perceptually*, by masking the stimulus details that provoke the illusion (i.e., covering the arrows in the Müller-Lyer, masking the checkerboard pattern in the checker-shade illusion, looking away from the McGurk video, etc.). Given the mechanism I posit underlies delusions like Capgras and MSM, I am suggesting that there *is no way* to "get around" the illusion via perception—no way to mask it—hence, it cannot be disassembled within the modality via which it is provoked.[27] Hence, the belief gets a free pass to awareness, although it may be inert from there (as sub-systems get their own information directly from perceptual modules, and never access the output of this higher-level integration assembly). And the belief perseveres because a) it never gets replaced with the right sort of content, and b) *every time it is even considered, it gets reinforced.* Why? Because the re-consideration involves remembering, which as discussed in the previous chapter, will essentially *recreate the episode* as it was encoded—which means the experience will be re-experienced.[28] And the experience automatically generates the "stranger" belief (or the "impostor" belief), and that belief is automatically endorsed. So it's a cul de sac. The delusional belief is stuck. The bell is rung and cannot be unrung.

8.3 Limitations of the modular analysis

I think this sort of modular analysis is very promising for explaining and modeling monothematic delusion. *Polythematic* delusions, however, may be more difficult to model this way, though my suspicion is that they *could be.* My own initial connection of monothematic delusion to modularity is because of the direct surface feature associations between such delusions and cross-modal *illusions*, which I have argued are certainly the result of modular breakdowns (or manipulations). Polythematic delusions are a different story: they are not circumscribed in the same strict ways. Persecutory delusions in particular are often highly *elaborated*, and provoke a great deal of behavioral response, as well as the formation of further elaborative inferences. Indeed, in the grip of a persecutory delusion, a person's entire cognitive system can become hostage to a single, central, false belief. This is the exact opposite of the sort of cul de sac of "belief" which results from localized modular misfires. The persecutory delusion is by definition *global*, not local.

Here's my speculative gesture toward a modular account, or at least a modular structure-constrained account, of persecutory delusion.[29] Let's take for granted some sort of anomalous perception that is (mis)interpreted, at some level of local processing, as a *bad social intention*—for instance, the possibility that a hearing deficit, when fused with lowered self-esteem, or a bias to over-attribute negative intention to others, could lead to an initial conclusion that one is being whispered about by others (McKay *et al.,* 2009). Maybe it's even as vague and as simple as that: some subsystem is putting out the message SOCIAL/INTENTIONAL THREAT. That information, given its immense salience, could conceivably engage all the sorts of heuristic processing and bias that have been discussed earlier in this dissertation, to marshal attentional and perceptual resources to the THREAT—essentially, all further perception and inference (and belief formation and revision) is *framed* under the assumption of THREAT. Bentall *et al.* (2001) suggest something similar to this, without drawing the obvious conclusion, in my opinion.

> In summary, it seems that a variety of perceptual and attentional factors may contribute toward the development or maintenance of persecutory ideation. Patients with persecutory delusions attend selectively to threat-related information, are excessively sensitive to others' expression of negative emotion, preferentially recall threat-related information, and give high estimates of the frequency with which those events occur to themselves and to others. However, they spend less time looking at threat-related information than ordinary people. (Bentall *et al.*, 2001: 1154)

I'm not sure that the comparisons to "ordinary" people here are actually apt. "Ordinary" people are subject to the exact same biases *when experiencing a salient threat*. The delusional person comes to the testing center already *besieged* by a perceptual/experiential environment that is broadcasting a (mistaken) elevated "threat" level. So yes, they attend preferentially (and less rigorously) to threat-related information, engage in confirmatory evidence evaluation regarding threats (and avoid disconfirmatory evidence) etc. This is the failsafe functioning of an overarching, virtually encapsulated system operating via heuristic algorithms. In addition, all the ongoing processing will be further cultured/corrupted by further memory-driven associations that essentially serve to *repeat* and thereby *reinforce the validity* of the (mistaken) threat detection alert. So the persecutory belief, once generated, is capable of *dominating* the global system—there is no rational "choice" involved in this. In that sense, I would agree with Maher (1974) that it's not really an *irrational process* at work. Indeed, all of our ostensibly rational systems are wired up precisely to promote perceived threats and license them to hijack processing. That's a good design principle, generally speaking, assuming underlying threat detection systems are operating normally, and that the threats detected, for the most part, are *real.* But a single corrupted threat detection signal can wreak major havoc on a system designed specifically to *put threats first* in the processing pecking order. "Just because you're paranoid doesn't mean they're not really after you", after all. Better to fail-safe than fail-dangerous.[30]

The other major limitation of my account—relying on modular structural constraints, virtually encapsulated processing assemblies and heuristically-driven search and judgment algorithms—is that is has a certain hand-waving, "just so" quality, that in the case of persecutory delusions, for example, may make my theory entirely unfalsifiable. For monothematic delusions, like Capgras and MSM, I have suggested specific experiment designs that could help confirm or disconfirm the modular etiological hypothesis. For elaborated delusions, it's not as clear how to test the hypothesis. Perhaps a hypnotically induced paranoia could be induced in "ordinary" subjects to see if they suddenly exhibit all the same sorts of biased evidence evaluation and preferential attention characteristics, as described by Bentall *et al.,* above. That might be dispositive confirmation of my model—in which a single, vague SOCIAL THREAT belief can quickly overtake the global reasoning and belief revision systems of the "ordinary" person in the same fashion. If so, I would

suggest that such a finding could support the more general, modularly constrained, massively parallel local processing account I have defended for belief revision in general. Recall my "bag of hammers" metaphor from §6.3: there I argued that our encapsulated, heuristically-programmed reasoning system(s) are like hammers in the sense that *to a person with a hammer, everything is a nail.* Our deliberative faculties and reasoning "tools" take the problems we face in a dynamic complex environment and *literally* transform them into something they can handle, at the representational level. This is very much the case for the paranoid person: once you've got your *threat* hammer out, everything suddenly looks like a threat.

8.4 Wrapping up

In this chapter, I have tried to show the connection between my account of belief revision and the phenomena of delusional belief. Delusions are exactly this sort of "bell that can't be unrung"—a species of unrevisable belief—that I have argued are inevitable products of a system designed for computationally feasible belief revision, which necessitates trading off precision for tractability. I have offered a specific argument that at least some forms of monothematic delusion are the result of localized deficits to cross-modal perceptual integration modules, and are thus directly analogous to the sorts of cross-modal illusions discussed in chapters 3-4 of this dissertation. I have also argued that this account of delusion as an integrative modular misfire can help resolve some live disputes in the delusion literature, specifically insofar as my account favors a one-factor, doxastic endorsement account of monothematic delusion.

Notes for chapter 8

[1] I admit this is an oversimplification, I do it in the interest of space.

[2] Are delusions no different than any old regular false belief state? Obviously not—delusions will be a subset of false belief in general. I'll explain this in more detail later, but for now let me say simply that delusions are a sort of false belief—the ways in which they diverge from "regular" (non-delusional) false belief does not require a new status designation. "Belief" will do just fine, on my account. *Cf.* Mele (1987; 2001; 2009) for useful discussions of how to cash out borderline delusional states, such as extreme instances of self-deception predicated on false belief states.

[3] See Frith (1987; 2012; also Hirjack & Fuchs, 2010; Jones & Fernyhough, 2007; Fu & McGuire, 2003) for more discussion of loss of agency and delusions of alien control in schizophrenia. *Cf.* Stephens & Graham (2000) for an overview on self-consciousness and alien voices.

[4] Though, to be fair, they train their fire equally on endorsement accounts for Capgras, which they say "require an explanation of the deficit causing the experience that a familiar person (or object) has been replaced . . . Without such an explanation, 'expression' accounts have no explanatory power" (Fine *et al.,* 2005: 148).

[5] Note that Gerrans (2009) disputes this reading of the Capgras delusion. He claims that it's not plausible that the experience can contain the "impostor" content, as that is essentially a claim about numerical and qualitative identity (i.e., "the person before me is qualitatively but not numerically identical to mother"), yet, as Bayne & Fernandez (2009) explains it, Gerrans objects that "the affective response is downstream from numerical identification rather than prior to it" (17).

[6] Wait, you might object: delusion is fundamentally different than illusion because you *can* realize your illusions aren't real and *believe* otherwise, whereas delusional subjects often cannot be disabused of their delusional beliefs. However, you only give up the *illusional belief* when you have *experienced its illusory nature*, directly, within the sensory modality within which that illusion arose. You don't take anyone's word for it that the M-L lines are actually equal, you need to *see* that. Same thing with Capgras. The problem, as we'll see below, is the underlying perceptual deficit in Capgras may make it *impossible* to "see".

[7] Note that if the impairment is to *observation*, then we arguably aren't necessarily talking about a second deficit, at least in cases where the delusional belief is generated *via observation*, i.e., perceptually. That's not a deficit in observation *on top* of the perceptual deficit—they are coextensive in that sense. Below, I will argue for a one-factor model for such perception-generated delusions—their maintenance is due exclusively to the fact that observational checking procedures simply regenerate the delusion in a closed loop. See Stone and Young (1997) for more discussion on the tension between "observational adequacy" and delusional beliefs that arise from perceptual deficits.

[8] It will be an endorsement account—also one-factor and doxastic, for the record, though that should be apparent from my comments in the section above.

[9] I will have more to say about how my account diverges from Pacherie's below, in the section dedicated to the Capgras delusion.

[10] Note, it can even extend to pets, and in some cases, to familiar *objects.* Abed & Fewtrell (1990) report the case of a woman (Mrs. S.) who "expressed the belief that a number of familiar objects had been replaced by near-identical duplicates" (1990: 915). *Cf.* Villarejo *et al.* (2011) for further examples, including a case where a man believed his belt and other personal belongings had been replaced by impostors.

[11] I will *not*, in the end, agree that there is an "explanation" stage here—except in the sense that unconscious, modular integration processes arrive at this "conclusion". I use "explanation" here loosely, in the short term, just to explain the delusion and the standard etiological story.

[12] Coltheart (2005) takes on this challenge, and attempts to defend the explanationist view:

> There is no reason to believe that everyone who suffers a disconnection between the face recognition system and the autonomic nervous system does come up with the same hypothesis… What seems to be true is that people with this hypothesis have this disconnection; that does not imply that all people with this disconnection come up with this hypothesis. But even if that were so, what other hypotheses, apart from the hypothesis of brain damage, are observationally adequate? (Coltheart, 2005: 155-56)

I don't find this convincing, especially since the "brain damage" explanation is not only *more* plausible, but usually *actually is suggested* to the patient. So Coltheart needs to be able to explain why the impostor "explanation" is not only uniform, but also not the most obvious one, and not the one supported by testimony from others. That sounds less and less like an explanation, and more and more like an endorsement of an automated "hypothesis" generation—one that admits no alternatives.

[13] Note that this model can also account for the double dissociation between Capgras and prosopagnosia: if the integrator receives a positive output from covert recognition, and a negative from overt face recognition, then that is fused accordingly: conscious awareness receives the report "unknown person" but all the lower level

systems tied into covert recognition directly will treat the person as familiar, hence SCR levels will register familiarity, etc.

[14] But wait, you might think – that actually *is* a crazy conclusion. Well, it's only crazy after reflection. As an *initial* conclusion, it's not crazy at all. It actually sort of makes sense. And here's an evolutionary argument for why it's adaptive: I would go ahead and simply assume that the covert system is more evolutionarily ancient (it's certainly much deeper below consciousness—much more System 1 than System 2, in that sense). If overt recognition evolved later, and integration even later, then it makes sense that early stages of that evolutionary process may have *benefitted* from an "impostor detector" – not because there were lots of impostors, but because overt recognition simply wasn't that good yet. And like every adapted system that we've discussed in this dissertation: it's always adaptive to be programmed to fail-safe rather than fail-dangerous. It's safer to incorrectly think a familiar is a stranger than the opposite. And as for it being crazy—the McGurk effect is kind of crazy too. Hearing imaginary /da/ even after it's proven to you it's not there? Crazy.

[15] It's the same for the Muller-Lyer illusion, for that matter. You only *know* that what you "see" is false once it's been proven *visually* to you (by masking the arrowheads, or providing a reference ruler). Also in the checker-shade illusion—you will not *believe* the squares are the same shade just because someone *tells you they are.* It wouldn't even be enough, I presume, to subject the squares to a spectrometer to "prove" it. You won't *believe* they are the same until you *see* that—by masking the picture to leave only the two squares.

[16] Recall above the finding of Hasher, Goldstein, Toppino (1977) report that mere *repetition* of information enhances perceived validity.

[17] Hirstein & Ramachandran (1997) describe an interesting "treatment" for Capgras involving their patient DS. The patient had come to believe his father had been replaced by an impostor, and after repeated attempts to disabuse DS of this had failed, the father instead tried to "play along" with the delusion

> He walked into his son's room one day and announced, ' the man who you have been with all these days is an imposter—he isn't really your father. I have sent him away to China. I am your real father—it's so good to see you son.' DS's delusion seemed to abate slightly after this 'treatment,'... Yet during a subsequent interview a week later DS had reverted to his original delusion, claiming that the imposter had returned. Also, his father told us in confidence that although DS had accepted him now as his father. (Hirstein & Ramachandran, 1997: 439)

I think this anecdote supports the argument that sufferers of Capgras are caught in a "vicious epistemic circle"—the *playing along* with the delusion allowed for a brief moment of it lifting, but DS's perceptual experiences with his father would not have been changed, and hence the delusion was regenerated. Note the important aspect in this case is that the delusional belief was *not* overridden: the father in fact supported the belief that the *previous* interactions had been with an impostor. All the father managed to do with his lie was temporarily block the *present* belief from forming, he was presenting himself as someone *different* than the impostor. But in the end, the experience of him *as* impostor overwhelms DS, and is endorsed. He is still an impostor, to DS. (Worse, he has know given some support to the impostor hypothesis... I'm not sure this is a good outcome!)

[18] Note that some researchers have reported cases of Capgras that *did* motivate behaviour (Christodoulou, 1986; Buchanan & Wessely, 2004), though these seem to be a minority of cases.

[19] This would be a good place to explain where my modular account of the Capgras delusion diverges from Pacherie's (2009) account, mentioned in the previous section. One area in which my account is quite different from hers is regarding the order of processing that goes on in (complete) recognition, integrating both covert and overt systems, Pacherie suggests that:

> Although this may go beyond the sense in which Fodor intended the notion of domain-specificity, the affective system that generates the sense of familiarity may still be considered domain-specific insofar as it takes as its inputs specific types of descriptors, such as face recognition units, voice recognition units, and other very fine-grained recognition units yielded by earlier perceptual analysis processes (2009: 20).

I think the analysis of familiarity is not quite right, however. I think what I will call *familiarity detection* is not a function that takes places *after* other, overt recognition mechanisms (face, voice, etc.) as Pacherie suggests, but rather is *prior to*, and *feeds* the integration function, as I have argued above. Recall the familiarity/recognition dissociation in the chapter on memory: implicit memories, encoded in the familiarity

system, are tied directly into many motivational and affective and behaviour guiding systems, beneath consciousness, in ways that recognition system memories are not.

Another area of difference between my and Pacherie's account is in regards to the *maintenance* of the (endorsed) delusional belief—why is it not overridden? We both agree on one fundamental point, that

> the reason why the Capgras patients fail to dismiss their delusional beliefs is not that they fail to use these checking procedures. Rather, it happens that these procedures fail to yield disconfirming evidence. For them to give solid grounds to reject the belief, the damaged module would have to be intact. The Capgras patient is not epistemically incompetent; rather, in a way, he is the victim of a vicious epistemic circle. (Pacherie, 2009: 120)

Pacherie doesn't tie the modularity thesis directly to the fact that the checking procedures fail to yield disconfirming evidence. She focuses, instead, on the general vagaries of background knowledge as evidence, and the fact that since, in Capgras, those *familiar* to the person are not to be trusted removes testimonial evidence from the evidence set (though it would usually be the best way to disabuse someone of a false belief). My argument is that *no* checking procedure can work except to re-experience the "impostor" in a way that s/he *is not perceived as an impostor*. But this will be impossible in this case, as the perceptual integration device is defective, and there is no "work-around" as in illusions: no way to disassemble the delusion without the sensory modality from which it is generated.

[20] Breen *et al.,* (2001) report that of the two patents with MSM delusions that they followed, one of them exhibited some facial recognition deficits, but the other did not. Clearly in the second case, facial recognition can't be the genesis of the MSM delusion. *Cf.* Villarejo *et al.* (2011) for more case studies of MSM.

[21] Connors *et al.,* defending a two-factor model, suggest the hypnotic state itself is the second factor in MSM:

> Interestingly, within hypnosis, the factor 1 alone suggestion was just as successful in generating the delusion as the combined Factor 1 and Factor 2 suggestion. This suggests that hypnosis, which itself is known to disrupt belief evaluation, can act as Factor 2 in this analogue. (2012: 18).

I disagree with this conclusion, again highlighting the generalizability argument: if hypnosis is the second factor here, then people under hypnosis should have defective belief formation system *across the board*, but that doesn't seem to be the case. Indeed, the fact that the delusion can be incited in healthy subjects, *without* prompting the specific explanation that it's a stranger, suggests that there is only one step to the delusion—the blocking of facial recognition (via hypnosis in this case) is enough to corrupt the sensory integration in a way that *immediately* ends up with the content-rich belief *stranger in the mirror*.

[22] Failsafe design again. And, of course, mirrors aren't really part of the natural environment in which our friend/foe detection systems evolved. Reflections would be, true—so an ability to recognize one's reflection could be useful. Of course, the reflections in question would have been much less clear and complete and much more ambiguous than what we get from mirrors in the modern world, so a cross-checking system would be adaptive.

[23] Let me be perfectly clear on this point, however: I am not arguing for a specific module adapted for this purpose, in the way that many of the other modules we have looked at *could* be. What I imagining here is a "virtual" module, one that can be functionally described, and which piggy-backs on other processes, not particularly concerned with mirrored-selves. But it is "modular" in the sense of meeting all the relevant criteria, *modulo* "non-assembly", though I have argued in chapter 4 that this criterion is unnecessarily strict anyhow.

[24] An alternate example would be when on a train, looking out the window, and the train next to you begins to move. In that case, since *you* are not moving your *body*, there is no easy way to determine which train is actually in motion, without another objective reference. You can't "frame" the perceptual scene via motor control information. Your perception will be ambiguous.

[25] Here's a puzzle for my account though: why *don't* they treat the video image as an *other*. For that matter, why don't we *all* do that? According to my account, the ID system should detect intentionality of anyone on a TV screen. How does *that* get inhibited? Well, I would suggest that there must be other systems that integrate with the visual perceptions ins a way that inhibits it—I can only speculate, but I would guess that the 2D nature on the screen "tells" the visual system something about the non-veridicality of what's in view. A mirror image, on the other hand, has perceptible depth (i.e., you can focus on deeper and shallower parts of the reflection),

whereas the TV screen has a set focal depth. Perhaps this alone is enough to account for the ID system not "insisting" on an intentional other *onscreen.* Of course, some deeply embedded System 1 processes *do* treat onscreen others as *actual others*—which is why we can respond emotionally to movies, be afraid, ascribe intentions (*cf.* Heider & Simmel, again) and even jump out of our seats (especially in 3D movies). So even though generally we are screen-knowledgeable enough to shut down some inferences, many of our (dumber) sub-systems proceed to take onscreen action as *actual* action, worthy of reflexive response.

[26] We know this because we use the mirror-recognition task to test for *self-awareness* (Gallup, 1970)—putting an ink spot on a child's forehead, then placing her before the mirror: does she reach for the spot in the mirror, or own her own head? If it's the latter, she obviously recognizes herself in the mirror. Human children generally exhibit the ability to pass this sort of test by 18-24 months of age. Many of our primate cousins are also capable of demonstrating awareness/understanding of their own reflection using this test (Gallup, 1970; 1982). So far, what this tells us is that mirrored-self recognition is an ability with a clearly established ontogenetic path.

[27] In the hypnosis study, Connors *et al.* (2012) note that the most successful "challenge" to break the delusion was to stand in front of the mirror with the subject and work it out logically, in real time, by "counting how many people are in the mirror" and then comparing that to how many people were in the room. This lets the subject *experience* the "reality" that it is her in the mirror after all. But even so, the response tends to be "I guess that is me....". The drive to believe otherwise is very strong.

[28] Recall Roediger (2000: 72), from chapter 7, above: "Retrieving is like perceiving for a sentient observer" … "due to the pre-eminence of retrieval processes, it may be possible to have a full-blown experience of remembering an event when the specific event was never encoded or stored".

[29] I won't be defending it—I'll leave that to a future endeavor.

[30] It's possible that not all delusions that fit into the "persecutory" schema necessarily involve the perception of THREAT at a level sufficient to provoke massive affective response and hijack the system. One recently coined delusion is the "Truman Show delusion" (Gold & Gold, 2012). In this delusion, patients "developed the delusional belief that they were the ''star'' of a reality television show secretly broadcasting their daily life, much like the main character in Peter Weir's film *The Truman Show*" (Gold & Gold 2012: 456). It's not very clear why this delusion of celebrity would necessarily provoke a *threatened* response in the way I have discussed regarding persecutory delusions in general—indeed, many people would be *thrilled* to be the star of a reality show—though the "secret" conspiratorial nature of the show in the Truman Show delusion is paranoia-infused. This type of delusion may be hard to model on an account such as the one I have presented. Similarly, any sort of *grandiose* delusion, of which the Truman Show delusion seems to overlap somewhat, may be hard to model for similar reasons.

Conclusion

My central aim in this dissertation has been to trace out an account of the cognitive architecture that human rationality, belief, and belief revision requires—not what it *normatively* requires, insofar as how cognition would have to be structured in order to meet the normative demands of rationality, but, rather, what actual belief revision *descriptively* requires, or appears to require. Given the interlocked constraints of the *frame problem* and the *finitary predicament*, I have argued that the only viable strategy for modeling belief revision will rely on massively parallel modular processing, aided by heuristic search and judgment procedures, and heavily circumscribed by the limits of recall in a fundamentally associative memory retrieval process. The payoff of a system of belief revision that is mediated entirely by subdoxastic modular functioning is that it is *tractable*—it can actually *work* and get its work *done*—whereas, a system that operates according to the way philosophers tend to talk about belief revision *cannot* work. The price of a system of belief revision that is mediated entirely by subdoxastic modular functioning is that it is *error-prone*— it will have systematic patterns of breakdown, and will be technically incapable of meeting the globally coherent, holistic principles of belief maintenance typically demanded by norms of rationality. The upshot of this is that we probably should accept a much more deflationary understanding of those norms and principles.

The standard argument for an account such as mine is generally a defensive one, played out on terms set by Jerry Fodor: he insists belief is isotropic and Quinean, whereas modules, due to informational encapsulation, are incapable of effecting isotropic or Quinean processes. My argument has followed two strands: on one, I have made the standard defensive move and attempted to show how a massively modular system, run on heuristic and associative processes, *can*, in the end, *approximate* the Quinean ideal when it comes to belief revision. The other strand of my argument is one that I believe is perhaps less common: I have attacked the notion that belief is isotropic and Quinean in the sense Fodor (and many others) take for granted. This attack is predicated on the fact that any belief revision process will necessitate numerous *recall* steps (i.e., we will need to retrieve from

memory various items, including other beliefs, in order to check for consistency, assess evidence, etc.). I have employed various arguments regarding conceptual and memory storage and retrieval, as well as a number of empirical studies on memory distortion and directed forgetting, which seem to show very clearly that recall is anisotropic: what is retrieved in any given act of recall will be shaped by the retrieval context, in ways that we can often neither control nor be conscious of. These limits of recall compromise the Quinean holism of our belief revision practices, but also make them possible, in terms of tractability.

Finally, my account predicts that a normally functioning belief revision system, if it is modular as I describe, will result in a number of unrevisable beliefs—bells that cannot be unrung. My conclusion is that the price of a system that can tractably revise belief *at all* is that it will *fail* to revise *all* belief. I have attempted to show that this prediction is borne out in the empirical literature on memory, false belief, and delusion. In the final chapter, I make a novel argument that certain monothematic delusional syndromes are directly related to the modular underpinnings of belief revision. This, in turn, helps to resolve many ongoing disputes in the literature on delusion regarding the doxastic status and etiology of certain delusions, and could help open up a new avenue for research into potential treatment.

Allow me to recap briefly, chapter by chapter, the main points of my argument in this dissertation. In chapter one, I began with a discussion of Quine & Ullian's *Web of Belief*, which served as a paradigm example of a standard normative discussion of belief revision practices which highlights coherence and conservatism as the primary virtues. Through a discussion of various phenomena of pervasive inconsistency, perseverant false belief, and misinformation effects that are difficult to notice and sometimes impossible to remedy, I have tried to show that the normative ideal of belief revision—in which one systematically weeds out inconsistencies in one's belief set—is largely beyond human capacity.

In chapter two, I further explored these limits, allying myself with Cherniak's diagnosis of the 'finitary predicament' of human cognition—the limited time and resources we have to devote to the seemingly insurmountable computational tasks that holistic belief revision and inferential thought demand. I expanded on his concerns by introducing what is known in computing and cognitive science as the *frame problem:* the question of how any system can *frame* a potentially infinite task ahead of time in order to make it tractable. Given

the task of belief revision, the frame problem poses the question of how much evidence needs to be considered before (dis)confirming any given belief. If we assume a Quinean stance on belief revision, the answer to that question seems to be *all the relevant evidence*—but this clearly courts a combinatorial explosion, insofar as determining what's relevant is (in turn) an unframed task. As a result, something needs to give.

In chapter three, I outlined the move to invoke modularity as the standard strategy to skirt the frame problem and account for the finitary predicament, at least with regard to sensory perception. I mounted a defense of perceptual modularity against examples of so-called cognitive penetration that I believe is somewhat novel. Rather than trying to save the modular account from evidence that suggests violations of encapsulation, I have argued that all the standardly discussed cases of cognitive penetration, rather than speaking against the modularity of perceptual systems, actually serve as evidence for further layers of modular processing, at the level of sensory integration. The *uniformity* and *robustness* of instances of purported cognitive penetration suggest the systematic breakdown patterns of a modular interface systems that run cross-modal error-correction and interpolative data-smoothing subroutines. If cognitive penetration were truly taking place, such effects should be *remediable*, and should be expected to *differ* in form from subject to subject. The fact that they are neither speaks to their being the byproduct of further, integrative modular systems.

In chapter four, I expanded on the notion of integrative modularity to include other sorts of higher-level modular assemblies: modules built out of modular subcomponents. Attempts to resolve the frame problem by invoking this sort of massive modularity thesis are not new, and I have highlighted a number of such attempts along the way, borrowing elements from a host of thinkers to aid in the elaboration of my own account. I think that where I have perhaps advanced the ball forward somewhat on the question is in the defense of assembled modules against Fodorian objections. For Fodor, assembled modules are a violation of the very definition of modularity, but I have tried to make a plausible case for Fodor's concerns being misplaced in this regard, and that the condition of "non-assembly" is too strict to capture what Fodor is actually concerned with—viz., blocking the sort of cognitive penetration that is taken as violative of the spirit of modules. I believe we can safely remove the prohibition against modular assembly and still preserve the core elements that make modularity useful: namely, the computational tractability that is gained via

encapsulation and domain-specificity. I examined a number of specific proposals for assembled modules that seem *prima facie* plausible in order to defend the idea in principle.

In chapter 5, I continued with my defense against Fodorian objections, attempting to show how Fodor's own theory of concept acquisition arguably requires massively modular processing to effect the sort of conceptual locking and filing his account proposes. Furthermore, I argued that the nature of Fodor's theory of concept acquisition and filing points the way towards resolving the frame problem: concepts self-organize in a highly compartmentalized way as an artifact of the manner in which they are acquired. As a result, later searching can run on exclusively associative processing—something Fodor claims will never work, but I believe I have shown *can* work quite effectively, even on his own theoretical terms. I further elaborated on this by bringing on board insights from Endel Tulving's work on memory retrieval to show that the empirical evidence regarding how we succeed (and fail) to recall items from storage suggests exactly the highly compartmentalized and associative filing that I have discussed and which gives us a path around the frame problem. I conclude with a point about the anisotropic, non-Quinean nature of recall from memory, arguing that so long as belief revision involves some steps of recall (which it seems plausible to suggest *all* instances of belief revision will involve), then belief revision will be *de facto* anisotropic and non-Quinean after all: it is simply not the case that every belief can in principle be brought to bear on every other belief, as Fodor claims. The limits of recall heavily circumscribe the processes of belief revision and deliberation.

Of course, the impulse to insist that belief revision *ought* to be Quinean is strong and well-founded. In chapter 6, I attempt to show how a massively parallel modular cognitive architecture can support processing that *approximates* Quinean belief revision in a satisfactory way. The key elements necessary to such a system are a global workspace and a set of nested heuristic algorithms and associative processes which can manage entry into that workspace, and invoke halting procedures on judgment procedures. I have turned to work by Baars and Jackendoff to support the global workspace hypothesis, and used the *Jeopardy*-winning A.I. Watson as a prime example of how a system can mimic human reasoning in a tractable fashion using a global workspace/blackboard architecture, and a few simple heuristics. I also examined a number of findings from the heuristics & biases research program that shed light on the sorts of heuristics that appear to be inherent in human

reasoning, and I have attempted to show how they are sufficient to approximate the holistic reasoning and belief revision practices we are concerned with explaining. I conclude this chapter by noting that heuristics—often referred to metaphorically as a adaptive cognitive "toolkit"—might be better described as a cognitive "bag of hammers," in the sense that they are simple tools which not only work effectively, but literally transform the problem space into tractable formats. To a brain with a hammer, every problem looks like a nail.

In the final two chapters, I have attempted to illustrate how numerous findings in psychology support the sort of cognitive architecture my account proposes, and bear out a number of predictions that I highlight at the end of chapter 6 regarding predictable patterns of breakdown in the belief revision process. In chapter 7, I look to research on memory distortion and perseverant false belief to illustrate how these phenomena have all the telltale signs of being byproducts of modular functioning, just as I have described in previous chapters. Additionally, research on how to direct people to *forget* confirms one of the major predictions of my account: there will be conditions under which certain things can never, properly speaking, be forgotten—certain bells cannot be unrung—and analogously, certain beliefs cannot be unbelieved, thus undermining standard normative accounts of belief revision. This predictable breakdown is the price paid for tractability and for skirting the frame problem. In chapter eight, I made the connection between my account of belief revision and the phenomena of delusional belief. Delusions are exactly the sort of "bell that can't be unrung"—a species of unrevisable belief—that I have argued are inevitable products of a system designed for computationally feasible belief revision, which necessitates trading off precision for tractability. I offered a specific argument that at least some forms of monothematic delusion are the result of localized deficits to cross-modal perceptual integration modules, and are thus directly analogous to the sorts of cross-modal illusions discussed in chapters 3-4. I also argued that this account of delusion as an integrative modular misfire can help resolve some live disputes in the delusion literature, specifically insofar as my account favors a one-factor, doxastic endorsement account of monothematic delusion.

Clearly there are many further issues and questions to pursue on the topics of belief revision and how it can work within a massively modular architecture. Many of the claims and arguments I have put forward here leave my account exposed to possible

counterexample. Additionally, a number of elements of the account only suggest the *prima facie* plausibility of massively parallel modular assemblies, but are certainly not dispositive regarding the existence of any particular such modular assemblies. Further research could perhaps confirm some of the various modular integration devices I have discussed. My hope is that some of the proposals and arguments contained in this dissertation might be useful in provoking further fruitful examinations of these questions in cognitive science, and in helping forge links between research programs in philosophy and psychology on the question of belief revision which may be highly mutually relevant, but not previously brought to bear on one another.

Works cited

Abed, R.T., Fewtrell, W.D. (1990). Delusional misidentification of familiar inanimate objects. A rare variant of Capgras syndrome. *The British Journal of Psychiatry*, *157*(6), 915-917.

Adelson, E.H. (2005). The checker-shadow illusion. MIT. Retrieved online, June, 2014. <http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html>

Allik, J. (2000). Available and Accessible Information in Memory and Vision. *Memory, consciousness, and the brain: the Tallinn conference*. Philadelphia, PA: Psychology Press, 7-17.

Amador, X.F., David, A.S. (2004). *Insight and psychosis* (2nd ed.). Oxford: Oxford University Press.

Amador,X.F., Kronengold. H. (2004) Description and Meaning of Insight in Psychosis. *Insight and psychosis* (2nd ed. pp 15-32). Oxford: Oxford University Press.

Anderson, C., Lepper, M., & Ross, L. (1980). Perseverance of social theories: the role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*(6), 1037-49.

Anderson, J.R. (1982). Acquisition of Cognitive Skill. *Psychological Review, 89,* 369-406.

Anderson, J.R., Bower, G.H. (1973). *Human Associative Memory.* Washington DC: Winston

Andrews, K. (2005). Chimpanzee theory of mind: Looking in all the wrong places?. *Mind & language*, *20*(5), 521-536.

Arkes, H.R., Boehm, L., Xu, G. (1991). Determinants of Judged Validity. *Journal of Experimental Social Psychology, 27,* 576-605.

Arkes, H.R., Hackett, C., Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making, 2*, 81-94.

Ariely, D. (2008). *Predictably irrational: the hidden forces that shape our decisions*. New York, NY: Harper.

Aspeitia, A.A.B., Eraña, Á., & Stainton, R. (2010). The contribution of domain specificity in the highly modular mind. *Minds and Machines 20*(1), 19-27.

Baars, B.J. (1988). *A cognitive theory of consciousness*. Cambridge: Cambridge University Press.

——— (1997). *In the theater of consciousness: the workspace of the mind*. New York: Oxford University Press.

Bach, K. (1999). The myth of conventional implicature. *Linguistics and philosophy*, *22*(4), 327-366.

Bach, M. (2014). Visual Phenomena and Optical Illusions. <http://www.michaelbach.de/ot/>.

Bach, N. (2011). A Comparison between the IBM Watson DEEPQA and Statistical Machine Translation. *Carnegie Mellon Language Technologies Institute,*

Baddeley, A.D. (1986). *Working memory*. Oxford: Oxford University Press.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The Adapted mind: evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Barnier, A. J., Cox, R. E., Connors, M., Langdon, R., & Coltheart, M. (2010). A Stranger in the Looking Glass: Developing and challenging a hypnotic mirrored-self misidentification delusion. *International Journal of Clinical and Experimental Hypnosis*, *59*(1), 1-26.

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge, Mass.: MIT Press.

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

Barrett, H.C. (2005). Enzymatic computation and cognitive modularity. *Mind and Language*, *20*, 259-87.

Barrett, H.C., Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological review 113*(3), 628-647.

Barsalou, L.W. (1992). *Cognitive psychology: an overview for cognitive scientists*. Hillsdale, N.J.: L. Erlbaum Associates.

——— (1999). Perceptual symbol systems. *Behavioural and Brain Sciences*, *22*, 577-660.

——— (2009). Simulation, situated conceptualization, and prediction. *philosophical transactions of the royal society*, *364*, 1281-89.

Bartlett, F.C. (1932) *Remembering.* Cambridge, Mass.: Cambridge University Press.

Basden, B. H., Basden, D. R., Gargano, G. (1993). Directed forgetting in implicit and explicit memory tests: a comparison of methods. *Journal of Experimental Psychology*, *19*(3), 603-616.

Bayes, T., Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton. *Philosophical Transactions of the Royal Society of London 53*: 370–418.

Bayne, T., Fernandez, J. (2009). *Delusion and self-deception: affective and motivational influences on belief*

formation. New York: Psychology Press.
Bayne, T., Pacherie, E. (2004). Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry, & Psychology*, *11*(1), 1-11.
——— (2005). In defence of the doxastic conception of delusions. *Mind & Language*, *20*(2), 163-188
Begg, I. Armour, V., Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science, 17,* 199-214.
Bentall, R. P., Kinderman, P., & Kaney, S. (1994). The self, attributional processes and abnormal beliefs: towards a model of persecutory delusions. *Behavioural Research Therapy*, *32*(3), 331-41.
Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N., & Kinderman, P. (2001). Persecutory delusions: A review and theoretical integration. *Clinical Psychology Review*, *21*(8), 1143-92.
Bickle, J. (2009). *The Oxford handbook of philosophy and neuroscience*. Oxford: Oxford University Press.
Bjork, E., Bjork, R. (2003). Intentional Forgetting can Increase, not Decrease, Residual Influences of To-Be-Forgotten Information. *Journal of Experimental Psychology: Learning Memory and Cognition*, *29*(4), 524-31.
Bjork, R. (1998). Intentional forgetting in perspective: comments, conjectures, and some directed remembering. *Intentional forgetting: interdisciplinary approaches* (pp. 453-82). Mahwah, N.J.: L. Erlbaum Associates.
——— (1989). Retrieval inhibition as an adaptive mechanism in human memory. In *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (eds. H.L. Roediger, F.I.M. Craik, pp.309-330). Hillsdale, NJ: Erlbaum.
Bjork, R., Woodward, A.E. (1973). Directed forgetting of individual words in free recall. *Journal of Experimental Psychology, 99,* 22-27.
Blakemore, S., Frith, C. (2003). Disorders of self-monitoring and symptoms of schizophrenia. *The self in neuroscience and psychiatry* (pp. 407-424). Cambridge, UK: Cambridge University Press.
Bloom, L. (1970). *Language Development: Form and function in emerging grammars*. Cambridge MA: MIT Press.
Bogacz R., Brown, M.W., Giraud-Carrier, C. (1999). High capacity neural networks for familiarity discrimination. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)* (Vol. 2, pp. 773-778).
——— (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, *10*(1), 5-23.
Bogdan, R.J. (1986). *Belief: form, content, and function*. Oxford: Clarendon Press.
Bortolotti, L. (2009). Delusion. *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/delusion/>.
——— (2010). *Delusions and other irrational beliefs*. Oxford: Oxford University Press.
Bota, R.G., Munro, J.S., Ricci, W.F., & Bota, D.A. (2006). The Dynamics of Insight in the Prodrome of Schizophrenia. *CNS Spectr*, *11*(5), 355-362.
Bowerman, M. (1982). 11. Reorganizational processes in lexical and syntactic development. *Language acquisition: The state of the art*, 319.
Braun-Latour, K.A., Latour, M.S., Pickrell, J.E., & Loftus, E. (2004). How and when advertising can influence memory for consumer experience. *journal of advertising*, *33*(4), 7-25.
Breen, N., Caine, D., Coltheart, M., Hendy, J., & Roberts, C. (2000). Towards an understanding of delusions of misidentification: four case studies. *Pathologies of belief* (pp. 75-110). Oxford: Blackwell.
Breen, N., Caine, D., Coltheart, M. (2001). Mirrored-self Misidentification: Two Cases of Focal Onset Dementia. *Neurocase*, *7*, 239-254.
Broome, J. (2007). Is Rationality Normative? *Disputatio, 2(*23), 161-178.
Broome, M.R., Bortolotti, L. (2009). *Psychiatry as cognitive neuroscience: philosophical perspectives*. Oxford: Oxford University Press.
Brown, M.W., Warburton, E.C. (2006). Associations and dissociations in recognition memory systems. In *Handbook of Binding and Memory* (eds. Zimmer, Mecklinger, Lindenberger), 413-444.
Buchanan, A., Reed, A., Wessely, S., Garety, P., Taylor, P., & Grubin, D. (1993). Acting on delusions. II: The phenomenological correlates of acting on delusions.. *British Journal of Psychiatry*, *163*, 77-81.
Buchanan, A., & Wessely, S. (2004). Delusions, actions and insight. *Insight and psychosis* (2nd ed., pp. 241-268). Oxford: Oxford University Press.
Buckwalter, W., Rose, D., Turri, J. (2013). Belief through thick and thin. *Noûs*. 1-23.
Buridan, John (1989). *Questions on Aristotle's de Anima* (in *John Buridan's Philosophy of Mind: an edition*

and translation of Book III, ed. and trans. J. Zupko). Ph.D. thesis: Cornell University.
Burnston, D., Cohen, J.L. (2014). Perceptual Integration, Modularity, and Cognitive Penetration. *Cognitive Influences on Perception.* Oxford: OUP. Forthcoming.
Byrne, A., Hilbert, D.R. (1987). *Readings on Color: The Philosophy of Color, vol. I.* Cambridge, MA: MIT Press.
Cain, M. J. (2002). *Fodor: language, mind, and philosophy*. Cambridge, UK: Polity.
Callebaut, W., & Gutman, D. (2005). *Modularity: understanding the development and evolution of natural complex systems*. Cambridge, Mass.: MIT Press.
Capgras, J., Reboul-Lachaux, J. (1994). L'illusion des 'sosies' dans un delire systematise chronique. . *History of Psychiatry*, *5*, 119-33.
Carruthers, P. (2003) "On Fodor's Problem". *Mind and Language*, 18(5): pp. 502-523.
——— (2006a). *The architecture of the mind: massive modularity and the flexibility of thought*. Oxford: Clarendon Press.
——— (2006b). Simple heuristics meet massive modularity. In *Carruthers, Laurence & Stich* (2006), 181-196.
——— (2006c). The case for massively modular theories of mind. In R. Stainton (ed.) *Contemporary Debates in Cognitive Science,* 7-21.
Carruthers, P. Laurence, S., Stich, S. (Eds.). (2005). *The Innate Mind: Structure and contents*. Oxford: OUP.
——— (2006). *The Innate Mind: volume 2: culture and cognition* (Vol. 2). OUP.
Carruthers, P., & Smith, P.K. (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.
Ceci, S.J. (1995) False beliefs: some developmental and clinical perspectives. In *Memory* Distortion(ed. D. Schacter). Cambridge, MA: Harvard Univ. Press, pp. 91-125.
Chabris, C., Weinberger, A., Fontaine, A., & Simons, D. (2011). You do not talk about Fight Club if you do not notice Fight Club: Inattentional blindness for a simulated real-world assault. *i-Perception*, *2(2)*, 150-53.
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.
Cherniak, C. (1984) Computational Complexity and the Universal Acceptance of Logic. *Journal of Philosophy 81*(12), 739-758.
——— (1986) *Minimal Rationality.* Cambridge, MA: Bradford.
Christiansen, M.H., Chater, N. (2001). *Connectionist Psycholinguistics.* Westport: Ablex
Christodoulou G.N. (1986). *Delusional Misidentification Syndromes*. Karger, Basel.
Chomsky, N. (1959). A Review of B. F. Skinner's *Verbal Behavior*. *Language, 35*(1), 26-58
Churchland, P., Ramamchandran, V., Sejnowski, T. (1994). A Critique of Pure Vision. *Large-scale neuronal theories of the brain* (pp. 23-60). Cambridge, Mass.: MIT Press.
Clark, H.H., & Clark E.V. (1977) *Psychology and Language.* New York: Harcourt Brace.
Clarke, M. (2004). *Reconstructing Reason and Representation.* Cambridge: MIT Press
Cohen, J. (2009). *The Red and the Real.* Oxford: OUP.
Cohen, J.L. (1992). *An Essay on Belief and Acceptance.* Cambridge: Clarendon Press.
Coltheart, M. (2005). Conscious Experience and Delusional Belief. *Philosophical Psychology*, *12*(2), 153-7.
Coltheart, M., Davies, M. (2000). *Pathologies of belief*. Oxford: Blackwell.
Connors, M., Barnier, A. J., Coltheart, M., Cox, R. E., Langdon, R. (2012). Mirrored-self misidentification in the hypnosis laboratory: Recreating the delusion from its component factors. *Cognitive Neuropsychiatry*, *15*(2), 151-76.
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. *Mapping the mind: domain specificity in cognition and culture* (pp. 85-116). Cambridge: Cambridge University Press.
Cotard, J. (1880). *Du Délire hypocondriaque dans une forme grave de la mélancolie anxieuse, mémoire lu à la Société médico-psychologique dans la séance du 28 juin 1880, par M. le Dr Jules Cotard...* Donnaud.
Craik, F.I., Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of experimental Psychology: general*,*104*(3), 268.
Curd, M, Cover, J.A. (1998) *Philosophy of Science.* New York: WW Norton and Co.
Currie, G. (2000). Imagination, delusion and hallucinations. *Pathologies of belief* (pp. 167-182). Oxford: Blackwell.

Currie, G., Jureidini, J. (2001). Delusion, Rationality, Empathy: Commentary on Martin Davies et al. *Philosophy, Psychiatry, & Psychology*, *8*(2), 159-162.

Czigler, I. Winkler, I., eds. (2010). *Unconscious Memory Representation in Perception.* Amsterdam: John Benjamins Publishing Co.

David, A. S. (2004). The clinical importance of insight. *Insight and psychosis* (2nd ed., pp. 332-351). Oxford: Oxford University Press.

Davies, A.M., Davies, M., Ogden, J.A., Smithson, M., White, R.C. (2009). Cognitive and motivational factors in anosognosia. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 187-117). New York: Psychology Press.

Davies, A.M., Davies, M. (2009). Explaining pathologies of belief. *Psychiatry as cognitive neuroscience: philosophical perspectives* (pp. 285-323). Oxford: Oxford University Press.

Davies, G., Wright, D.B. (2010). *Current issues in applied memory research*. New York, NY: Psychology Press.

Davies, M. (2009). Delusion and motivationally biased belief: self-deception in the two factor framework. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 71-86). New York: Psychology Press.

Davies, M., Coltheart, M., Langdon, R., Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, *8*(2), 133-158.

Deaner, R.O., Shepherd, S.V., Platt, M.L. (2007). Familiarity accentuates gaze cuing in women but not men. *Biology Letters*, *3*(1), 65-68.

Dedrick, D. (2009). *Computation, cognition, and Pylyshyn*. Cambridge, Mass.: MIT Press.

Dehaene, S., Kerszberg, M., Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529-14534.

Dehaene, S., Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition, 79, 1–37.

Dehaene, S., Sergent, C., Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Science,

Delk, J.L. and Fillenbaum, S. (1965). ‘‘Differences in Perceived Colour as a Function of Characteristic Color,’’ *The American Journal of Psychology*, 78 (2): 290–93.

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

Descartes, R. (1641/1985). *The philosophical writings of Descartes* (Vol. 2). Cambridge University Press.

Devine, D., Clayton, L., Dunford, B., Seying, R., & Pryce, J. (2000). Jury Decision Making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy and Law*, *7*(3), 622-727.

Dretske, F. (1986). Misrepresentation. *Belief: form, content, and function* (pp. 17-36). Oxford: Clarendon Press.

Dreyfus, H. (1972). *What Computers Can't Do*, New York: MIT Press.

——— (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, Oxford, U.K.: Blackwell.

——— (1992). *What Computers* Still *Can't Do*, New York: MIT Press.

Egan, A. (2009). Imagination, delusion and self-deception. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 263-280). New York: Psychology Press.

Eich, E., Metcalfe, J. (1989). Mood dependent memory for internal versus external events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 443.

Ellis, H.D. (1996). Delusional misidentification of inanimate objects: A literature review and neuropsychological analysis of cognitive deficits in two cases. *Cognitive neuropsychiatry*, *1*(1), 27-40.

Ellis, H.D., & Lewis, M.B. (2001). Capgras delusion: a window on face recognition. *Trends in Cognitive Science*, *5*(4), 149-156.

Ellis, H.D., Young, A.W. (1990). Accounting for delusional misidentifications. *The British Journal of Psychiatry*, *157*(2), 239-248.

Ellis, H.D., de Pauw, K.W., Christodoulou, G.N., Papageorgiou, L., Milne, A.B., Joseph, A.B. (1993). Response to facial and non-facial stimuli presented tachistoscopically in either or both visual fields by patients with the Capgras delusion and paranoid schizophrenics. *Journal Neurol. Neurosurgery Psychiatry 56*, 215-219.

Ellis, H.D., Young, A.W., Quayle, A.H., De Pauw, K.W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London. Series B: Biological*

Sciences, *264*(1384), 1085-1092.
Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., Plunkett, K. (2001). *Rethinking Innateness: A connectionist perspective on development (5th ed.).* Cambridge: MIT Press.
Elman, J.L., McClelland, J.L. (1988). Cognitive Penetration of the Mechanisms of Perception: Compensation for Coarticulation of Lexically Restored Phonemes. *Journal of Memory and Language 27*, 143-165.
Elmes, D.G., Wilkinson, W.C. (1971). Cued forgetting in free recall: Grouping on the basis of relevance and category membership. *Journal of Experimental Psychology*, *87*(3), 438.
English, S. M., & Nielson, K. A. (2010). Reduction of the misinformation effect by arousal induced after learning. *Cognition*, *117*, 237-242.
Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic Why the adjustments are insufficient. *Psychological science*, *17*(4), 311-318.
Evans, J.S., Frankish, K. (2009). *In two minds: dual processes and beyond*. Oxford: Oxford University Press.
Fein, S., Morgan, S., Norton, M., & Sommers, S. (1997). Hype and Suspicion: The Effects of Pretrial Publicity, Race, and Suspicion on Juror's Verdicts. *Journal of Social Issues*, *53*(3), 487-502.
Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press
Festinger, L., Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, *58*(2), 203.
Festinger, L., Maccoby, N. (1964). On resistance to persuasive communications. *The Journal of Abnormal and Social Psychology*, *68*(4), 359.
Fine, C. (2006). *A mind of its own: how your brain distorts and deceives*. New York: W.W. Norton & Co..
Fine, C., Craigie, J., & Gold, I. (2005a). The Explanation Approach to Delusion. *Philosophy, Psychiatry, & Psychology*, *12*(2), 159-63.
Fine, C., Craigie, J., Gold, I. (2005b). Damned if You Do; Damned if You Don't: The Impasse in Cognitive Accounts of the Capgras Delusion. *Philosophy, Psychiatry, and Psychology* , *12*(2), 143-151.
Fine, C., Gardner, M., Craigie, J., Gold, I. (2007). Hopping, Skipping or Jumping to Conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry*, *12*(1), 46-77.
Flexser, A.J., Tulving, E. (1978). Retrieval independence in recognition and recall. *Psychological Review*, *85*(3), 153.
Ferrucci, D., Brown, E., Chu-Carroll, C., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J. Schlaefer, N., Welty, C. (2010). Building Watson: An overview of the DeepQA Project. *Association for the Advancement of Artificial Intelligence.* 59-79
Fish, S. (2011). What Did Watson the Computer Do? *New York Times, Opinionator,* Feb. 21, 2011, < http://opinionator.blogs.nytimes.com/2011/02/21/what-did-watson-the-computer-do/ >
Fodor, J.A. (1975). *The language of thought*. New York: Crowell.
——— (1983). *The modularity of mind: an essay on faculty psychology*. Cambridge, Mass.: MIT Press.
——— (1984). Observation reconsidered. *Philosophy of Science*, 23-43.
——— (1998). *Concepts: where cognitive science went wrong*. Oxford: Clarendon Press ;.
——— (2000). *The mind doesn't work that way: the scope and limits of computational psychology*. Cambridge, Mass.: MIT Press.
——— (2005) "Reply to Pinker's 'So How Does the Mind Work?'" *Mind & Language 20 (1),* pp. 25-32
——— (2008). *LOT 2: the language of thought revisited*. Oxford: Clarendon Press ;.
Fodor, J.A., Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition 58*(2), 253-270.
——— (2002). *The Compositionality Papers*. Oxford: Clarendon Press.
Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
——— (2009). Delusions: a two level framework. *Psychiatry as cognitive neuroscience: philosophical perspectives* (pp. 269-284). Oxford: Oxford University Press.
Frazier, L., Fodor, J.D. (1978). The Sausage machine: A new two-stage parsing model. *Cognition 6,* 291-325.
Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic bulletin & review*, *5*(3), 490-495.
Frith, C. (1987). The positive and negative symptoms of schizophrenia reflect impairments in the initiation and perception of action. *Psychological Medicine 134*, 225-235.
——— (2006). *Making up the mind: how the brain creates our mental world*. Oxford: Blackwell.
——— (2012). Explaining delusions of control: The comparator model 20years on. *Consciousness and cognition*, *21*(1), 52-54.
Fu, C., & McGuire, P. (2003). Hearing voices or hearing the self in disguise?. *The self in neuroscience and*

psychiatry (pp. 425-35). Cambridge, UK: Cambridge University Press.
Gallagher, S. (2009a). Delusional realities. *Psychiatry as cognitive neuroscience: philosophical perspectives* (pp. 245-267). Oxford: Oxford University Press.
——— (2009b). Delusional experience. *The Oxford handbook of philosophy and neuroscience* (pp. 513-21). Oxford: Oxford University Press.
——— (2005). *How the body shapes the mind*. Oxford: Clarendon Press.
Gallistel, C.R. (1990). *The organization of learning*. Cambridge, Mass.: MIT Press.
Gallup, G.G. (1970). Chimpanzees: self-recognition. *Science*, *167*(3914), 86-87.
——— (1982). Self-awareness and the emergence of mind in primates. *American Journal of Primatology*, *2*(3), 237-248.
Garety, P. (2004). Insight and delusion. *Insight and psychosis* (2nd ed., pp. 66-77). Oxford: Oxford University Press.
Garety, P., Freeman, D., Jolley, S., Dunn, G., Bebbington, P. E., Kuipers, E., et al. (2005). Reasoning, Emotions, and Delusional Conviction in Psychosis. *Journal of Abnormal Psychology*, *114*(3), 373-84.
Garety, P., Hemsley, D., Wessely, S. (1991). Reasoning in deluded schizophrenic and paranoid patients: biases in performance on a probabilistic inference task. *Journal of Nervous and Mental Disease*, *179*(4), 194-201.
Garfield, J.L., Peterson, C.C., Perry, T. (2001). Social cognition, language acquisition and the development of the theory of mind. *Mind & Language*,*16*(5), 494-541.
Gazzaniga, M., Miller, M. (2000). Testing Tulving: the split brain approach. *Memory, consciousness, and the brain: the Tallinn conference* (pp. 307-318). Philadelphia, PA: Psychology Press.
Gazzaniga, M. S., Bizzi, E. (1995). *The cognitive neurosciences*. Cambridge, Mass.: MIT Press.
Gendler, T.S. (2007). Self-deception as pretense. *Philosophical Perspectives*, *21*, 231-258.
——— (2008a). Alief in action (and reaction). *Mind and Language*, *23*(5), 552-85.
——— (2008b). Alief and belief. *Journal of Philosophy*, *105*(10), 634-63.
Gerken, L. A. (1994). 'Young children's representation of prosodic phonology: evidence from English-speaker's weak syllable omissions'. *Journal of Memory and Language 1:* 19-38.
Gerrans, P. (2000). Refining the explanation of Cotard's delusion. *Pathologies of belief* (pp. 111-122). Oxford: Blackwell.
——— (2009). From phenomenology to cognitive architecture and back. *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*, 127-138.
Gibson, J.J. (1979) *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin
Gigerenzer, G. (2000) *Adaptive Thinking: Rationality in the Real World*. New York: Oxford U. Press
——— (2011). *Heuristics: the foundations of adaptive behavior*. Oxford: Oxford University Press.
Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: the adaptive toolbox*. Cambridge, Mass.: MIT Press.
Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*(2), 107-119.
——— (2002). Inferential Correction. *Heuristics and biases: the psychology of intuitive judgment* (pp. 167-184). Cambridge, U.K.: Cambridge University Press.
Gilbert, D.T., Krull, D.S., Malone, P. S. (1990). Unbelieving the unbelievable: some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601-613.
Gilbert, D.T., Tafarodi, R.W., Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221-233.
Gilbert, D.T., & Wilson, T.D. (2009). Why the brain talks to itself: sources of error in emotional prediction. *Philosophical Transaction of The royal Society*, *364*, 1335-1341.
Gilchrist, A., Cowan, N. (2010) Conscious and unconscious aspects of working memory (In *Unconscious Memory Representation in Perception, ed.* Czigler & Winkler), 1-36.
Gilovich, T. (1991). *How we know what isn't so: the fallibility of human reason in everyday life*. New York, N.Y.: Free Press.
Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge, U.K.: Cambridge University Press.
Givón, T. (2005) *Context as Other Minds: The Pragmatics of Sociality, Cognition and Communication*. Amsterdam: John Benjamins Publishing Co.
Gleitman, L. R., and Gillette, J. (1995). 'The role of syntax in verb learning'. In *The Handbook of Child Language* (eds. Fletcher & MacWhinney). Cambridge: Blackwell.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of psychology*,*66*(3), 325-331.
Gold, I., Hohwy, J. (2000). Rationality and Schizophrenic Delusion. *Mind & Language*, *15*(1), 146-67.
Gold, J., Gold, I. (2012). The 'Truman Show' Delusion: Psychosis in the Global Village. *Cognitive Neuropsychiatry*, *1*. Retrieved June 5, 2012, from http://dx.doi.org/10.1080/13546805.2012.666113
Golding, J.M., & Long, D.L. (1998). There's more to intentional forgetting than directed forgetting: an integrative review. *Intentional forgetting: interdisciplinary approaches* (pp. 59-102). Mahwah, N.J.: L. Erlbaum Associates.
Golding, J.M., & MacLeod, C.M. (1998). *Intentional forgetting: interdisciplinary approaches*. Mahwah, N.J.: L. Erlbaum Associates.
Gangopadhyay, N., Madary, M., Spicer, F. (Eds.). (2010). *Perception, Action, and Consciousness: Sensorimotor dynamics and two visual systems*. Oxford University Press.
Gopnik, I., Gopnik, M. (1986). *From models to modules: studies in cognitive science from the McGill Workshops*. Norwood, N.J.: Ablex Pub. Corp..
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.
Gould, J. L. (1990). Honey bee cognition. *Cognition*, *37*(1), 83-103.
Gould, J. L., Gould, C. G. (1988). *The honey bee*. Scientific American Library.
Gould, S.J., Lewontin, R.C. (1979). The Spandrels of San Marco and the Panglossian Paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B, 205*(1161), 581-598.
Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing Inferences During Narrative Text Comprehension. *Psychological Review*, *101*(3), 371-395.
Graf, P., Mandler, G. (1984). Activation makes words more accessible, but not necessarily more retrievable. *Journal of Verbal Learning and Verbal Behavior*, *23*(5), 553-568.
Gregory, R.L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 279-296.
Grice, H.P. (1975). Logic and Conversation. In P. Cole and J. Morgan (eds.) Syntax and Semantics Volume 3: Speech Acts.
——— (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
Hadamard, J. (1923). *Lectures on the Cauchy problem in linear partial differential equations.* New Haven, CT: Yale University Press
Hannon, B., Craik, F.I. (2001). Encoding specificity revisited: The role of semantics. *Canadian Journal of Experimental Psychology*, *55*(3), 231.
Harman, G. (1986) *Change in View.* Cambridge, MA: MIT Press.
Hasher, L., Goldstein, D., Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107-112.
Hasher, L., Zacks, R.T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *American Psychologist*,*39*(12), 1372.
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.
Hirjak, D., & Fuchs, T. (2010). Delusions of Technical Alien Control: A Phenomenological Description of Three Cases. *Psychopathology*, *43*(2), 96-103.
Hirschfeld, L. A., Gelman, S. A. (1994). *Mapping the mind: domain specificity in cognition and culture*. Cambridge: Cambridge University Press.
Hirsh-Pasek, K., Golinkoff, R. M. (1996). *The Origins of Grammar.* Cambridge: MIT Press.
Hirstein, W. (2005). *Brain fiction: self-deception and the riddle of confabulation*. Cambridge, Mass.: MIT Press.
——— (2009a). Confabulations about people and their limbs, present or absent. *The Oxford handbook of philosophy and neuroscience* (pp. 473-512). Oxford: Oxford University Press.
——— (2009b). *Confabulation: views from neuroscience, psychiatry, psychology, and philosophy*. Oxford:

Oxford University Press.
Hirstein, W., Ramachandran, V.S. (1997). Capgras syndrome: a novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*(1380), 437-444.
Hodges, J.R., Patterson, K., Tyler, L.K. (1994). Loss of semantic memory: implications for the modularity of mind. *Cognitive Neuropsychology, 11*(5), 505-542
Hohwy, J., & Rosenberg, R. (2005). Unusual experiences, reality testing and delusions of alien control. *Mind & language*, *20*(2), 141-162.
Hood, B. M., Willen, J. D., & Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychological Science*, *9*(2), 131-134.
Horn, L. (1989). A natural history of negation. *IL: University of Chicago Press, Chicago*.
Hötting, K., Röder, B. (2004). Hearing cheats touch, but less in congenitally blind than in sighted individuals. *Psychological Science*, *15*(1), 60-64.
Hume, D. (1739/1978) *A Treatise of Human Nature, 2nd edition* (*eds*. L.A. Selby-Bigge & P.H. Nidditch) Oxford: OUP
Isbell, L.A. (2006). Snakes as agents of evolutionary change in primate brains. *Journal of Human Evolution 51,* 1-35.
Ito, T., Tiede, M., Ostry, D. J. (2009). Somotosensory function in speech perception. *PNAS*, *106*(4), 1245-48.
Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, Mass.: MIT Press.
Jackendoff, R. (1992). *Languages of the mind essays on mental representation*. Cambridge, Mass.: MIT Press.
Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford: Oxford University Press.
——— (2007). *Language, consciousness, culture essays on mental structure*. Cambridge, Mass.: MIT Press.
Jacob, P., Jeannerod, M. (2003). Ways of seeing: The scope and limits of visual cognition.
Jacoby, L.L., Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306.
Jacoby, L.L., Hollingshead, A. (1990). Toward a generate/recognize model of performance on direct and indirect tests of memory. *Journal of Memory and Language*, *29*(4), 433-454.
Johnson, H.M. (1998). Disregarding information in text. *Intentional forgetting: interdisciplinary approaches* (pp. 219-238). Mahwah, N.J.: L. Erlbaum Associates.
Johnson, H.M., Seifert, C.M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.
——— (1999). Modifying mental representations: Comprehending corrections. *The construction of mental representations during reading*, 303-318.
Johnson-Laird, P.N. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness.* Cambridge: Cambridge University Press
Jones, S.R., Fernyhough, C. (2007). Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and Cognition*, *16*, 391-99.
Kabanikhin, S.I. (2008). Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-posed Problems 16*, 317-357.
Kahneman, D., Tversky, A. (2000). *Choices, values, and frames*. New York: Russell sage Foundation
Kahneman, D., Fredrickson, B.L., Schreiber, C.A., Redelmeier, D.A. (1993). When more pain is preferred to less: Adding a better end.*Psychological science*, *4*(6), 401-405.
Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.
Kahneman, D., Knetsch, J.L., Thaler, R.H. (1991). Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *The Journal of Economic Perspectives 5*(1), 193-206.
Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*(4), 48-52.
——— (1979). *Organization in vision: Essays on Gestalt perception* (p. 1). New York: Praeger.
——— (1985). Seeing and thinking. *Acta Psychologica*, *59*(1), 23-33.
Karmiloff-Smith, A. (1979). Micro-and Macrodevelopmental Changes in Language Acquisition and Other Representational Systems*. *Cognitive Science*, *3*(2), 91-118.
——— (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition*, *23*(2), 95-147.
——— (1992). *Beyond modularity: a developmental perspective on cognitive science*. Cambridge, Mass.: MIT

Press.
Karmiloff, K., Karmiloff-Smith, A. (2001). *Pathways to Language.* Cambridge: Harvard University Press.
Kassin, S.M., Sommers, S.R. (1997). Inadmissible testimony, instructions to disregard, and the jury: Substantive versus procedural considerations. *Personality and Social Psychology Bulletin*, *23*, 1046-1054.
Keane, B.P., Lu, H., Papthomas, T.V., Silverstein, S.M., Kellman, P.J. (2012). Is interpolation cognitively encapsulated? Measuring the effects of belief on Kanizsa shape discrimination and illusory contour formation. *Cognition 123*(3), 404-418.
Kiersky, J. E. (2004). Insight, self-deception and psychosis in mood disorders. *Insight and psychosis* (2nd ed., pp. 91-106). Oxford: Oxford University Press.
Kinoshita, S. (2002). Feeling of familiarity. In *Metacognition* (pp. 79-90). Springer US.
Kircher, T., & David, A.S. (2003). *The self in neuroscience and psychiatry*. Cambridge, UK: Cambridge University Press.
Kiverstein, J. (2010). Sensorimotor knowledge and the content of experience. In *Perception, Action, and Consciousness: Sensorimotor dynamics and two visual systems (Ed.s* Gangopadhyay, N., Madary, M., Spicer, F.), 257-274
Koch, C., & Davis, J. L. (1994). *Large-scale neuronal theories of the brain*. Cambridge, Mass.: MIT Press.
Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago press.
Lakoff, G. (1987) *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press
Lakoff, G., Johnson, M. (1980) *Metaphors We Live By*. Chicago: University of Chicago Press
Lamb, T.D., Collin, S.P., Pugh, E.N. (2007). Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. *National Review of Neuroscience 8*(12), 960-76.
Laney, C., Loftus, E. (2010). Change blindness and eyewitness testimony. *Current issues in applied memory research* (pp. 142-60). New York, NY: Psychology Press.
Langdon, R. (2010). *Delusion and confabulation: a special issue of cognitive neuropsychiatry*. Hove: Psychology.
Langdon, R., & Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Pathologies of belief* (pp. 183-216). Oxford: Blackwell.
Levin, D. T., Simons, D. J., Angelone, B. L., & Chabris, C. F. (2002). Memory for centrally attended changing objects in an incidental real-world change detection paradigm. *British Journal of Psychology*, *93*(3), 289-302.
Liberman A.M., Mattingly I.G. (1985). The motor theory of speech perception revised. *Cognition 21*(1), 1–36.
Lillard, A. (1998). Ethnopsychologies: cultural variations in theories of mind. *Psychological bulletin*, *123*(1), 3.
Lipton, P. (2004). *Inference to the best explanation*. Psychology Press.
LoBue, V., DeLoache, J.S. (2008). Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science 19*(3), 284-89.
Locke, J. (1693). Letter to William Molyneux, 28 March, in *The Correspondence of John Locke* (9 vols.), E.S. de Beer (ed.), Oxford: Clarendon Press, 1979, vol. 4, no. 1620.
Loftus, E. F. (1974). Reconstructing memory: The incredible eyewitness. *Jurimetrics J.*, *15*, 188.
——— (2000). Remembering what never happened. *Memory, consciousness, and the brain: the Tallinn conference* (pp. 106-118). Philadelphia, PA: Psychology Press.
——— (2005). Planting misinformation in the human mind: a 30 year investigation of the malleability of memory. *Learning and Memory*, *12*, 361-66.
Lord, C.G., Lepper, M.R., Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of personality and social psychology*, *47*(6), 1231.
Lord, C.G., Ross, L., Lepper, M.R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.
Lutz, B.J. (2014). Coating on Willis Lower Skydeck's Ledge Crack Under Tourists. *NBC Chicago News.* Retreived May 30, 2014 <url: http://www.nbcchicago.com/news/local/chicago-willis-tower-sky-deck-ledge-crack-261079001.html>
Lycan, W. (1986). Tacit belief. *Belief: form, content, and function* (pp. 61-82). Oxford: Clarendon Press.
Mack, A., Rock, I. (1998). *Inattentional blindness*. The MIT Press.
MacLeod, C. M. (1998). Directed forgetting. *Intentional forgetting: interdisciplinary approaches* (pp. 1-58). Mahwah, N.J.: L. Erlbaum Associates.
MacLeod, C. M. (1999). The item and list methods of directed forgetting: test differences and the role of

demand characteristics. *Psychonomic Bulletin and Review*, *6*(1), 123-29.
Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research*, *84*(1), 24-62.
MacSwan, J. (2012). 13 Code-Switching and Grammatical Theory. *The Handbook of Bilingualism and Multilingualism*, 323.
Maher, B.A. (1974). Delusional thinking and perceptual disorder. *Journal of individual psychology 30*(1), 98-113.
——— (1988). Delusions as the product of normal cognitions. *Delusional beliefs* (pp. 333-336). New York: Wiley.
May, J. (2008) Man-machine. *James May's Big Ideas.* BBC. First broadcast in the UK, Sep. 15, 2008. Available online at <http://www.open.edu/openlearn/whats-on/ou-on-the-bbc-james-mays-big-ideas>
Marcus, G. (2009). *Kluge: The haphazard evolution of the human mind*. Houghton Mifflin Harcourt.
Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.
Marr, D., Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *204*(1156), 301-328.
Matthen 'Two Visual Systems and the Felling of Presence. In *Perception, Action, and Consciousness: Sensorimotor dynamics and two visual systems (Ed.s* Gangopadhyay, N., Madary, M., Spicer, F.), 107.
McClelland, J.L. (1987). The case for interactionism in language processing. *Artificial Intelligence and Psychology Project.* <http://repository.cmu.edu/psychology/426>
McClelland, J.L. (1995) Constructive Memory and Memory Distortions: A Parallel- Distributed Processing Approach. In *Memory Distortion* (ed. D. Schacter). Cambridge, MA: Harvard UNiv. Press, pp. 69-90.
McClelland, J.L., Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159.
McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General*, *114*(1), 1.
McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.*, *27*, 1-28.
McGilvray, J. (2002). 'Introduction for Cybereditions' of *Cartesian Linguistics*, by N. Chomsky. Christchurch: Cybereditions.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-48.
McKay, R., Dennett, D. C. (2009). The Evolution of Misbelief. *Behavioural and Brain Sciences*, *32*, 493-561.
McKay, R., Langdon, R., Coltheart, M. (2007). Models of misbelief: Integrating motivational and deficit theories of delusion. *Consciousness and Cognition*, *16*, 932-41.
——— (2005). "Sleights of Mind" delusions and self-deception. *Cognitive Neuropychiatry 10*(4), 305-326.
McKay, R., Cipolotti, L. (2007). Attributional style in a case of Cotard delusion. *Consciousness and cognition 16*(2), 349-359.
McNally, R.J., Lasko, N. B., Clancy, S. A., Macklin, M. L., Pitman, R. K., & Orr, S. P. (2004). Psychophysiological responding during script-driven imagery in people reporting alien abduction. *Psychological Science*, *15*(7), 493-97.
Mele, A. (1987). *Irrationality: an essay on akrasia, self-deception, and self-control*. New York: Oxford University Press.
——— (2001). *Self-deception unmasked*. Princeton, N.J.: Princeton University Press.
——— (2009). Self-deception and delusions. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 55-70). New York: Psychology Press.
Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and cognition 15*(3), 620-627.
Mercier, H., Sperber, D. (2009). Intuitive and Reflective Inferences. *In two minds: dual processes and beyond* (pp. 149-170). Oxford: Oxford University Press.
——— (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioural and Brain Sciences*, *34*, 57-111.
Metzger, W. (1934). Tiefenerscheinungen in optischen Bewegungsfeldern. *Psychol. Forsch., 20*, pp. 195‑206.
Meyers, L., & Fontana, W. (2005). Evolutionary Lock-in and the Origin of Modularity in RNA Structure. *Modularity: understanding the development and evolution of natural complex systems* (pp. 129-142). Cambridge, Mass.: MIT Press.

Milanovic, B. (2014). Global Income Inequality by the Numbers in History and Now: An Overview. *The World Bank Development Research Group, Policy Research Working Paper 6259.* Available online at: <http://elibrary.worldbank.org/doi/pdf/10.1596/1813-9450-6259>

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action* (Vol. 27).

Molyneux, W. (1693). Letter to John Locke, 2 March, in *The Correspondence of John Locke* (9 vols.), E.S. de Beer (ed.), Oxford: Clarendon Press, 1979, vol. 4, no. 1609.

Morewedge, C.K., Gilbert, D.T., Wilson, T.D. (2005). The least likely of times how remembering the past biases forecasts of the future. *Psychological Science*, *16*(8), 626-630.

Mundale, J., Gallagher, S. (2009). Delusional Experience. *The Oxford handbook of philosophy and neuroscience* (ed. J. Bickle). Oxford: Oxford University Press.

Nathaniel-James, D.A., Frith, C.D. (1996). Confabulation in Schizophrenia: Evidence of a new form? *Psychological Medicine 26(2)*, 391-400.

Niedenthal, P.M., Ric, F., Krauth-Gruber, S. (2002). Explaining emotion congruence and its absence in terms of perceptual simulation. Psychological Inquiry, 13, 80-83.

Nigro, G., Neisser, U. (1983). Point of view in personal memories. *Cognitive Psychology*, *15*(4), 467-482.

Nisbett, R.E., Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment.

Nisbett, R.E., Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, *84*(3), 231.

Öhman, A. (1993). Fear and anxiety as emotional phenomena: Clinical phenomenology, evolutionary perspectives, and information processing mechanisms. In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 511–536). New York: Guilford Press.

Öhman, A., Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological review*, *108*(3), 483.

Oltmanns, T. F., & Maher, B. A. (1988). *Delusional beliefs*. New York: Wiley.

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

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic bulletin & review*, *11*(6), 988-1010.

Pacherie, E. (2009). Perceptions, Emotions and Delusions. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 107-26). New York: Psychology Press.

Paller, K.A. (1990). Recall and stem-completion priming have different electrophysical correlates and are modified differentially by directed forgetting. *Journal of Experimental Psychology*, *16*(6), 1021-32.

——— (2006). Binding memory fragments together to form declarative memories depends on cross-cortical storage. (In *Handbook of Binding and Memory,* eds. Zimmer, Mecklinger, Lindenberger), 527-544.

Pea, R. (1980). The development of negation in early child language. *The social foundations of language and thought* (ed. D. Olson, pp. 156-186). New York: Norton

Pelli, D.G., Tillman, K.A. (2008). The uncrowded window of object recognition. *Nature neuroscience*, *11*(10), 1129-1135.

Peretz, I., Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, *6*(7), 688-708.

Pessoa, L., Thomson, E., Noe, A. (1998). Finding out about filling in: A guide to perceptual completion for visual science and philosophy of perception. *Behavioral and Brain Sciences 21*, 723-802.

Petty, R.E., Wegener, D.T. (1993) Flexible Correction Processes in Social Judgment: Correcting for Context-Induced Contrast. *Journal of Experimental Social Psychology 29(2)*, 137-165.

Pickel, K. (1995). Inducing Jurors to Disregard Inadmissable Evidence: A Legal Explanation Does Not Help. *Law and Human Behviour*, *19*(4), 407-24.

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

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

Pinker, S., Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1), 73-193.

Plunkett, K., Marchman, V. (1991). U-shaped learning and frequency effects in multi-layered perception: Implications for child language acquisition. *Cognition 38*(1), 43-102.

Poggio, T. (1981). Marr's computational approach to vision. *Trends in NeuroSciences 4*(10), 258-262.

——— (1985). Early Vision: from computational structure to algorithms and parallel hardware. *Computer Vision, Graphics, and Image Processing 31*, 139-155.

Poggio, T. & Koch, C. (1985). Ill-posed problems in early vision: from computational theory to analogue networks. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 226*(1244), 303-323.

Popper, Karl (1953/1998). Science: Conjectures and Refutations. In *Philosophy of Science* (ed. M. Curd & J.A. Cover) New York: W. W. Norton and Company, Inc.

Porter, S., Yuille, J.C., Lehman, D.R. (1999). The nature of real, implanted, and fabricated memories for emotional childhood events: implications for the recovered memory debate. *Law and human behavior*, *23*(5), 517.

Prinz, J.J. (2006). Is the mind really modular? In R. Stainton (ed) *Contemporary debates in cognitive science*, 22-36.

Putnam, H. (1988). *Representation and Reality*, Cambridge, Massachusetts: MIT Press.

Pylyshyn, Z.W. (1984). *Computation and cognition: toward a foundation for cognitive science*. Cambridge, Mass.: MIT Press.

——— (1987). *The Robot's dilemma: the frame problem in artificial intelligence*. Norwood, N.J.: Ablex.

——— (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3), 341–423.

Quine, W.V.O. (1951). Two Dogmas of Empiricism. *Philosophical Review 60*, 20-43.

Quine, W.V.O., Ullian, J.S. (1978) *The Web of Belief.* New York: McGraw-Hill.

Radden, J. (2011). *On delusion*. New York: Routledge.

Raftopoulos, A. (Ed.). (2005). *Cognitive penetrability of perception: attention, action, strategies, and bottom-up constraints*. Nova Publishers.

Ramachandran, V.S. (2011). *The tell-tale brain: a neuroscientist's quest for what makes us human*. New York: W.W. Norton.

Ramachandran, V.S., & Blakeslee, S. (1998). *Phantoms in the brain: probing the mysteries of the human mind*. New York: William Morrow.

Raymond, J.E., Shapiro, K.L., Arnell, K.M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of Experimental Psychology: Human perception and performance*, *18*(3), 849.

Reimer, M. (2009). Is the impostor hypothesis really so preposterous? Understanding the Capgras experience. *Philosophical Psychology*, *22*(6), 669-86.

Rey, G. (1988). Toward a computational account of akrasia and self-deception. *Perspectives on self-deception* (pp. 264-296). Berkeley: University of California Press.

Rice, H. J., Rubin, D. C. (2009). I can see it both ways: First-and third-person visual perspectives at retrieval. *Consciousness and cognition*, *18*(4), 877-890.

——— (2011). Remembering from any angle: the flexibility of visual perspective during retrieval. *Consciousness and cognition*, *20*(3), 568-577.

Richerson, P.J., Boyd, R. (2005). *Not By Genes Alone: How culture transformed human evolution.* Chicago: Univ. of Chicago Press

Ringach, D.L., Shapely, R. (1996). Spatial and temporal properties of illusory contours and amodal shape completion. *Vision Research 36*(19), 3037-50.

Roediger, H.L. (2000). Retrieval is the key process in understanding human memory. *Memory, consciousness, and the brain: the Tallinn conference* (pp. 52-75). Philadelphia, PA: Psychology Press.

Roediger, H.L., Jacoby, D., & McDermott, K.B. (1996). Misinformation effects in recall: creating false memories through repeated retrieval. *Journal of Memory and Language*, *35*, 300-318.

Ross, L., Lepper, M.R., Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32(5),* 880-892.

Rowlands (2005) The cognitive penetrability of perception. In Raftopoulos (ed.) *Cognitive penetrability of perception: attention, action, strategies, and bottom-up constraints*. Nova Publishers.

Rozin, P., Millman, L., Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology 50(4),* 703-712.

Rozin, P., Nemeroff, C. (2002). Sympathetic Magical Thinking: The Contagion and Similarity 'Heuristics'. *Heuristics and biases: the psychology of intuitive judgment* (pp. 201-216). Cambridge, U.K.: Cambridge University Press.

Rozin, P., Tuorila, H. (1993). Simultaneous and temporal contextual influences on food acceptance. *Food Quality and Preference 4(1)*, 11-20.

Ryan, L., & Eich, E. (2000). Mood dependence and implicit memory. *Memory, consciousness, and the brain: the Tallinn conference* (pp. 91-105). Philadelphia, PA: Psychology Press.

Ryle, G. (1949/2009). *The concept of mind*. Routledge.

Sackheim, H.A. (2004) The Meaning of Insight. *Insight and psychosis* (2nd ed. pp 3-14). Oxford: Oxford University Press.

Sackheim, H.A., Wegner, A.Z. (1986) Attributional patterns in depression and euthymia. *Archives of General Psychiatry, 4,* 553-560.

Sakagami, Y., Watanabe, R., Aoyama, C., Matsunaga, S., Higaki, N., Fujimura, K. (2002). The intelligent ASIMO: System overview and integration. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on* (Vol. 3, pp. 2478-2483). IEEE.

Sams, M., Mottonen, R., & Sihvonen, T. (2005). Seeing and hearing others and oneself talk. *Cognitive Brain Research*, *23*, 429-35.

Samuels, R. (2005). The Complexity of Cognition: tractability arguments for massive modularity. In Carruthers, Laurence, Stich (eds.) *The Innate Mind: Structure and contents*, 107-121.

——— (2006) Is the Human Mind Massively Modular? In R. Stainton (ed.) *Contemporary Debates in Cognitive Science, pp. 37-56.*

Schacter, D. (1995) *Memory Distortion.* Cambridge, MA: Harvard Univ. Press

Schacter, D. L., Eich, J. E., & Tulving, E. (1978). Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 721-743.

Searle, J.R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(03), 417-424.

Schkade, D.A., Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction.*Psychological Science*, *9*(5), 340-346.

Schul, Y., Burnstein, E. (1985). When discounting fails: Conditions under which individuals use discredited information in making a judgment. *Journal of Personality and Social Psychology, 49,* 894-903.

Schwitzgebel, E. (2010). Acting Contrary to Our Professed Beliefs, or The Gulf Between Occurent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly, 91(4)*, 531-553.

Seeck, M., Michel, C.M., Mainwaring, N., Cosgrove, R., Blume, H., Ives, J., Schomer, D.L. (1997). Evidence for rapid face recognition from human scalp and intracranial electrodes. *Neuroreport*, *8*(12), 2749-2754.

Segal, G. (1996). The modularity of theory of mind. *Theories of theories of mind* (pp. 141-157). Cambridge: Cambridge University Press.

Semon, R. (1904). Die Mneme als erhaltendes Prinzip im Wechsel des organischen Geschehens. Leipzig: Wilhelm Engelmann.

Seifert, C. M. (2002). The continued influence of misinformation in memory: what makes a correction effective?. *Psychology of Learning and Motivation: Advances in Research and Theory*, *41*, 265-92.

Sekuler, R., Sekuler, A., Lau, R. (1997). Sound alters visual motion perception. *Nature 385*, p.308.

Sergent, C., & Dehaene, S. (2004). Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink.*Psychological Science*, *15*(11), 720-728.

Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1997). The attentional blink. *Trends in cognitive sciences*, *1*(8), 291-296.

Shermer, M. (2008). Why People Believe Weird Things. *TED.* Available online at <https://www.youtube.com/watch?v=8T_jwq9ph8k>

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

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

Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129-138.

——— (1982). *Models of bounded rationality: Empirically grounded economic reason* (Vol. 3). MIT press.

Simons, D.J., Chabris, C.F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception-London 28(9),* 1059-1074.

Sinha, P., Balas, B., Ostrovsky, Y., Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, *94*(11), 1948-1962.

Sloman, S.A. (2002). Two Systems of Reasoning. *Heuristics and biases: the psychology of intuitive judgment* (pp. 379-396). Cambridge, U.K.: Cambridge University Press.

Soriano, M., Jimenez, J., Roman, P., & Bajo, M. (2009). Intentional Inhibition in Memory and Hallucinations: Directed Forgetting and Updating. *Neuropsychology*, *23*(1), 61-70.

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

Sperber, D., & Wilson, D. (2002). Pragmatics, Modularity and Mind-Reading. *Mind and Language*, *17*, 3-23.
——— (1995). *Relevance: communication and cognition*. Cambridge, MA: Blackwell Publishers.
Spezio, M. L., & Adolphs, R. (2009). Emotion, cognition, and belief: findings from cognitive neuroscience. *Delusion and self-deception: affective and motivational influences on belief formation* (pp. 87-138). New York: Psychology Press.
Spinoza, B. (1677/1994) *Ethics* (trans. Edwin Curley). Princeton: Princeton Univ. Press
Stainton, R. (2006). *Contemporary Debates in Cognitive Science.* Malden, MA: Blackwell Publishing.
Stainton, R., Viger, C. (2000). Review Essay: Jerry Fodor, *Concepts: Where Cognitive Science Went Wrong. Synthese 128,* 131-151.
Stanovich, K., West, R.F. (2000) individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences 23,* 645-726.
Stanovich, K.E.; West, R.F.; Toplak, M.E. (2013). "Myside Bias, Rational Thinking, and Intelligence". *Current Directions in Psychological Science 22* (4): 259–264.
Stephens, G. L., & Graham, G. (2000). *When self-consciousness breaks alien voices and inserted thoughts*. Cambridge, MA: MIT Press.
Stich, S.P. (1986). Are belief predicates systematically ambiguous? *Belief: form, content, and function* (pp. 119-148). Oxford: Clarendon Press.
Stone, T., Young, A.W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, *12*(3-4), 327-364.
Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology 18*(6): 643–662.
Sutton, J. (2012). Memory. *The Stanford Encyclopedia of Philosophy* (Ed. Edward N. Zalta) URL = <http://plato.stanford.edu/archives/win2012/entries/memory/>
Taylor, S.E., Fiske, S.T. (1975). Point of view and perceptions of causality. *Journal of Personality and Social Psychology, 32(3),* 439-445.
Thompson, C. (2010) "Smarter Than You Think: What is IBM's Watson?". *New York Times Magazine.* Accessed online < http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html >
Thomson, W., & MacLeod, F. (1998). "The jury will disregard...": a brief guide to inadmissible evidence. *Intentional forgetting: interdisciplinary approaches* (pp. 435-52). Mahwah, N.J.: L. Erlbaum Associates.
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
Tousignant, J.P., Hall, D., & Loftus, E. (1986). Discrepancy detection and vulnerability to misleading post-event information. *Memory and Cognition*, *14*(4), 329-338.
Tulving, E. (1972). Episodic and semantic memory 1. *Organization of Memory. London: Academic*, *381*, e402.
——— (1983). *Elements of episodic memory*. Oxford: Clarendon Press.
——— (1989). Remembering and knowing the past. *American Scientist*, 361-367.
——— (1991). Concepts of human memory. *Memory: Organization and locus of change*, 3-32.
——— (2000). *Memory, consciousness, and the brain: the Tallinn conference*. Philadelphia, PA: Psychology Press.
Tulving, E., Donaldson, W., & Bower, G.H. (1972). *Organization of memory*. New York: Academic Press.
Tulving, E., Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review 80(5),* 352-373.
Tversky, A., Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science 185(4157)*, 1124-1131.
Villarejo, A., Martin, V. P., Moreno-Ramos, T., Camacho-Salas, A., Porta-Etessam, J., & Bermejo-Pareja, F. (2011). Mirrored-self misidentification in a patient without dementia: evidence for right hemisphere and bifrontal damage. *Neurocase*, *17*(3), 276-84.
Vinden, P.G. (1996). Junin Quechua children's understanding of mind. *Child Development*, *67*(4), 1707-1716.
——— (1999). Children's understanding of mind and emotion: A multi-culture study. *Cognition & Emotion*, *13*(1), 19-48.
Vroomen, J. de Gelder, B. (2004). Sound Enhances Visual Perception: Cross-modal Effects of Auditory Organization on Vision. *Journal of Experimental Psychology 26,* 1583-1590.
Walster, E. (1967). 'Second guessing' important events. *Human Relations 20(3),* 239-249.

Wang, M. (2010). Implicit memory, anaesthesia and sedation. *Current issues in applied memory research* (pp. 165-184). New York, NY: Psychology Press.

Warren, R.M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392-393.

Warren, R.M., Warren, R.P. (1970). *Auditory illusions and confusions*. WH Freeman.

Warrington, E.K., Weiskrantz, L. (1974). The effect of prior learning on subsequent retention in amnesic patients. *Neuropsychologia*, *12*(4), 419-428.

——— (1978). Further analysis of the prior learning effect in amnesic patients. *Neuropsychologia*, *16*(2), 169-177.

Wason, P.C. (1960), "On the failure to eliminate hypotheses in a conceptual task", *Quarterly Journal of Experimental Psychology*(Psychology Press) *12* (3): 129–140,

——— (1968), "Reasoning about a rule", *Quarterly Journal of Experimental Psychology* (Psychology Press) *20* (3): 273–28,

Wegener, D.T. Petty, R.E. (1995). Flexible correction processes in social judgment: the role of naïve theories in corrections for perceived bias. *Journal of Personality and Social Psychology 68(1)*, 36.

Wegner, D.M. (1994). Ironic Processes of Mental Control. *Psychological Review 101(1),* 34-52.

——— (1997). When the Antidote is the Poison: Ironic Mental Control Processes. *Psychological Science*, *8*(3), 148-50.

Wegner, D.M., Coulton, G.F., Wenzlaff, R. (1985). The transparency of denial: Briefing in the debriefing paradigm. *Journal of Personality and Social Psychology 49(2),* 338-346Weinstein, E.A., Lyerly, O.G., Cole, M., Ozer, M.N. (1966). Meaning in Jargon Aphasia. *Cortex 2(2),* 165-187.

Wegner, D.M., Shortt, J.W., Blake, A.W., Page, M.S. (1990). The suppression of exciting thoughts. *Journal of Personality and Social Psychology, 58*, 409‑418

Weiskrantz, L. (1986). *Blindsight a case study and implications*. Oxford: Clarendon.

——— (1997). *Consciousness lost and found: a neuropsychological exploration*. Oxford: Oxford University Press.

Wenzlaff, R., & Wegner, D.M. (2000). Thought suppression. *Annual Review of Psychology*, *51*, 59-91.

Wilkes, A.L., Leatherbarrow, M. (1988). Editing episodic memory following the identification of error. *The Quarterly Journal of Experimental Psychology 40(2)*, 361-387.

Wilkes, A.L., Reynolds, D.J. (1999). On Certain Limitations Accompanying Readers' Interpretations of Corrections in Episodic Text. *The Quarterly Journal of Experimental Psychology 52(1)*, 165-183.

Wilson, T.D., Centerbar, D.B., Brekke, N. (2002). Mental Contamination and the Debiasing Problem. *Heuristics and biases: the psychology of intuitive judgment* (pp. 185-200). Cambridge, U.K.: Cambridge University Press.

Wistrich, A., Guthrie, C., & Rachlinski, J. (2005). Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding.. *Cornell Law Faculty Publications*, *4*, 1251-1345.

Wolf, S., Montgomery, D.A. (1977). Effects of inadmissible evidence and level of judicial admonishment to disregard on the judgment of mock jurors, *Journal of Applied Social Psychology*, **7**: 205–219.

Wyer, R.S., Budesheim, T.L. (1987). Person memory and judgments: the impact of information that one is told to disregard. *Journal of Personality and Social Psychology*, *53*(1), 14-29.

Wyer, R.S., Unverzagt, W.H. (1985). Effects of instructions to disregard information on its subsequent recall and use in making judgments. *Journal of Personality and Social Psychology 48(3),* 533.

Young, G. (2008). Restating the role of phenomenal experience in the formation and maintenance of the Capgras delusion. *Phenomenology and Cognitive Science*, *7*, 177-189.

Young, A.W., Leafhead, K.M. (1996). Betwixt life and death: Case studies of the Cotard delusion. *Method in Madness: Case studies in cognitive neuropsychiatry* (1996), 147-171.

Zimmer, H.D., Mecklinger, A., & Lindenberger, U. (2006). *Handbook of binding and memory: perspectives from cognitive neuroscience*. Oxford: Oxford University Press.

Zupko, J. (1989). *John Buridan's Philosophy of Mind: an edition and translation of Book III of His 'Questions on Aristotle's De Anima' (Third Redaction), with Commentary and Critical and Interpretative Essays*. Doctoral dissertation, Cornell University.