The Architecture of Belief:
An Essay on the Unbearable Automaticity of Believing

Eric Mandelbaum

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy

Chapel Hill
2010

Abstract

Eric Mandelbaum
The Architecture of Belief:
An Essay on the Unbearable Automaticity of Believing

My dissertation maintains that people cannot contemplate a proposition without believing that proposition. I present evidence and arguments that, contrary to popular opinion, we cannot withhold assent from any proposition we happen to consider. A model of belief fixation is sketched and used to explain hitherto disparate, recalcitrant, and somewhat mysterious psychological phenomena and philosophical paradoxes. Toward this end I also contend that our intuitive understanding of the workings of introspection is mistaken. In particular, I argue that propositional attitudes are beyond the grasp of our introspective capacities. We learn about our beliefs from observing our behavior, not from introspecting our stock beliefs.

The model of belief fixation offered in the dissertation poses a novel dilemma for theories of rationality. One might have thought that the ability to contemplate ideas while withholding assent is a necessary condition on rationality. In short, it seems that rational creatures shouldn't just form their beliefs based on whatever they happen to think. However, it seems that we are creatures that automatically and reflexively form our beliefs based on

whatever propositions we happen to consider. Thus, either the rational requirement that states that we must have evidence for our beliefs must be jettisoned or we must accept the conclusion that we are necessarily irrational.

Dedication

To all the people who have helped me make it through the morass that is graduate school. I'm not sure whether such completion was necessarily a good thing, but I do know that without the people mentioned in the acknowledgements, there would be no document sitting before you.

Acknowledgements

Although this document bears my name alone, the essay is not the product of my solitary efforts. One needs more than bread and water to survive and throughout graduate school I've been lucky enough to have many people help me hone my ideas and maintain my sanity (such as it is). Perhaps I haven't yet learned all the lessons my teachers have tried to impart to me, but surely without their attempts I'd have little to offer the world at large. The monumental task of explaining who I need to thank and why is one that I can barely begin to undertake, never mind complete, under the time pressure that I currently face. However, I will now attempt to give an incomplete list of those who deeply deserve my gratitude. I apologize in advance for all the people who have helped me who haven't been named here.

Fred Dretske has taught me that one can be lighthearted and caring, while still having a bulldog of an analytic mind. He embodies a rare combination of grace, humor, and trenchant insight. When I have doubted my abilities, I have found confidence in the fact that he believed in my ability to make a difference as a philosopher. Being able to interact with someone like Dretske justifies spending my twenties in graduate school in the south. Bill Lycan has counseled me through my many mishaps, both philosophically and personally. He has somehow always managed to engage me with humor while posing devastating and motivating objections, all the while doing so with the most awe-inspiring mustache I could imagine. Joshua Knobe broadened my horizons and afforded me opportunities that I could have never imagined, never mind achieved, without him. The patience, kindness, and intellectual originality he displays are unrivaled. Michael Corrado taught me how one can

succeed in academia without having to compromise one's morals. Perhaps this is the lesson I wish to have internalized the most. Ram Neta has seen this project through, giving me patient advice and convincing me that perhaps I am not the most anxiety ridden person on the planet. Ram, I wish us both the gift of solace. And to Jerry Fodor without whom…I would have been screamed at many fewer times.

My family has also been a bastion of stability throughout the tumult of graduate school. My sister has constantly cheered me on throughout this journey and my brother has acted as a best friend, counselor, and coach, always there to motivate and congratulate at every turn. My parents have also supported me however they could. My mother taught me the compassion and patience needed to make it through graduate school and my father's skeptical questioning and prodding motivated me to always work a bit harder. Any time I thought the load before me was overwhelming I'd think about my grandparents, who are no longer with us, and their struggles to survive the Holocaust and soviet work camps. The strength and grace they imbued taught me more than any classroom could ever hope to.

My friends have deeply shaped the intellectual aspects of this document. One learns more about their subject from their graduate student colleagues than from any other group. I have been lucky enough to surround myself with a dedicated, hilarious, fulfilling small group of impressive minds. In particular, I'd like to thank three people who have traveled this journey with me. Mark Phelan has been my lifeline to sanity in a sea of absurdity. Mark was always there to a give me an outlet to vent or to go out and grab a drink and just talk. Dave Ripley, my coauthor and dear friend, has shaped the content of this document more than perhaps any other person. Without his help, the section of negation would not exist. He has

also helped teach me what true intellectual adventurousness is. Being around Dave is one of the most inspiring things an academic could do. Lastly, Bryce Huebner has stuck with me as

the most loyal advocate one could ever hope for. He's read nearly everything I've ever written, heard all my crackpot ideas, and yet never ceases to heave voluminous and insightful comments. All three men are brothers to me. If I could change one thing about my life, it would be for all of us to still live in the same place.

Lastly, without my sweet, beautiful, insightful partner I'd probably be waking up in random sewers. My love for and appreciation of Molly Balikov is beyond compare. Molly has taught me how to be a moral human being with a full life. She is the reason I wake up smiling and go to sleep easy. She is my partner in this grand adventure and I know of no greater joy than simply basking in her company.

Any success I've achieved is a product of those people I have mentioned. The effort they've put into my life is truly astounding.

Preface

The only thing I find more impressive than human intelligence is the capacity for such impressive creatures to display such utter stupidity. One needn't look hard to find unbridled competence bound with unrivaled folly. I find this marriage most blatant in graduate school where supremely ingenuous folks often engage in such frivolous ventures. Of course, I too am not beyond such reproach. However, I find censure more oft-putting than ignorance so I'm not interested in playing the blame game. Instead, I want to understand a small part of the human condition. This essay is a meditation on a small aspect of our incompetence: our inclination to believe without warrant.

Table of Contents

List of Figures

Figures

Chapter 1: Introduction: Speculative Psychology Redux

I'm going out to clean the pasture spring;
I'll only stop to rake the leaves away
(And wait to watch the water clear, I may)
I sha'n't be gone long.—You come too.

I'm going out to fetch the little calf
That's standing by its mother. It's so young,
It totters when she licks it with her tongue.
I sha'n't be gone long.—You come too.

("The Pasture", Frost 1915/1995)

This essay has a fairly straightforward thesis: that one cannot entertain a proposition without believing it. There are caveats and qualifiers ahead, with some sub-conclusions, but the goal is clear enough: to convince the reader that the central thesis is if not true, then at least the best current theory of belief fixation. In this introductory chapter, I won't argue for the theory at all. Instead I will discuss some methodological presuppositions and meta-theoretical considerations in favor of actually engaging in the project before me. Since most of the project has already been engaged, such a venture is mootish, but I find little shame in railing against absurdity, so let's get to it.

My project is one that is concerned with empirical theory construction. This type of project has previously been labeled 'speculative psychology' (Fodor 1975), but I don't see that there's too much to be gained by labeling it. Perhaps what I'm doing doesn't quite count as psychology because it isn't an experimental science, and perhaps it isn't philosophy because the theories I'm trying to construct are empirical theories, theories about human

cognition. I'd prefer to think that my endeavor is both philosophical and psychological, but if you'd prefer you can just think of it as cognitive science, the great catch-all phrase for the discipline that isn't ever exactly sure what it's studying.[1] But regardless of how one tags the work, the goal is clear enough. What we want to do is to understand a bit about the human condition, specifically about how the mind works. I have no pretensions that I'll actually find out how the mind works, but I'd be happy with settling for finding out how the mind doesn't work. At this point I'm reasonably confident about how a bit of it, belief fixation, doesn't work, though a bit less sure about how it does. But this is not the worst place to be, because the view of the mind that I think is wrong is utterly ubiquitous. Thus, if I'm right about how we've been wrong, I'd consider that progress. Because no one likes a critic I'll also offer positive proposal about how I think belief fixation (in part) works. I offer this not just to offer another theory, but because this one strikes me as a fruitful research program. Of course, this is cognitive science, so it's probable that nothing any of us say will actually be true. But if I don't just get on with it I might as well switch fields, and it's too late in the game for that so there's no use in kvetching.

In what follows I take myself to be giving an abductive argument. What I want to do is to specify a hypothesis that has wide-ranging consequences. Part of my argument will show how this proposal fits the extant data on belief acquisition. However, another part of my argument will be of the following form: if we assume the architecture I propose, than we can explain a lot more, seemingly unrelated and problematic data of different sorts (e.g. the fundamental attribution error, the 'mere possibilities' formulation of the confirmation bias, the anchoring and adjustment heuristic, the puzzle of 'fearing fictions', etc.). Thus, I take it

[1] Not that most disciplines do know what their studying. As Ryle pointed out, good theory generally precedes good meta-theory (Ryle 1949).

that my proposal's explanatory and unificatory power outside the realm of belief acquisition is also evidence for the proposal. As mentioned, I'd bet that my proposal shares a central feature of all cognitive architectural proposals: it will probably turn out to be false. However, I think that my proposal is closer to the truth than the other dominant proposals on offer. As such, even if it's not true it may be a noble lie.

****

This project is a small piece of a vision of the mind that I find appealing. The picture of the mind that is in the background is one that I think I share with many other philosophers and cognitive scientists. Its presuppositions are sometimes murky and the overall view is rough around the edges, but it is a provocative and surprisingly robust picture. Since I think it's the emerging consensus view of the mind and since it's within this backdrop that the following dissertation resides, I'd like to spend a few moments outlining some main features of what I take the emerging consensus of the mind to be.

Roughly speaking, regardless of what one's pet projects are, almost any area of study within cognition will use the distinction between automated and controlled processes. Sometimes this distinction manifests itself in the ever-so-unclear literature on Type 1 and Type 2 processes (for an example more or less drawn at random see Wilson et al. 2000); sometimes it can manifest itself as the difference between 'top down' and 'bottom up' processing stories; sometimes it can manifest itself as the difference between 'informationally encapsulated' and 'inferentially unencapsulated' mental processes (Fodor 1983), etc. This is not meant to be a comprehensive list of the ways the distinction can manifest itself, rather, the previous list is just meant to suggest that there is a real division between processes that we, at the person level, can *control* and one's that we can't.

Without much recent fanfare, the concept of control has proven to be a central one in cognitive theorizing. The concept is of the utmost importance in both cognitive science and philosophy. It not only serves to distinguish between (roughly) two different types of mental processes, but it also seems central to our overall sense of self (Wegner 2003), our psychological and physiological health (Langer 1975; Rodin and Langer 1977), our theories of freedom, responsibility, and blame (Rosen 2004), doxastic voluntarism (James 1896/1992; Audi 2008) etc. In short, getting clear on the types of control is a central and important project across both philosophy and cognitive science. If I'm lucky, then this essay will be, in part, a small step toward such clarity.

It may seem that such clarity really isn't needed because identifying which processes are controlled and which are automatic is easy enough without philosophical exposition. But I suspect that this is not so: just as almost everyone accepts a division between automatic and controlled processes, so I suspect that in the background everyone knows that there is a continuum over which some processes can be seen as more or less controlled than others. There are some clear cases of cognitive processes that truly do seem to be automatic and others that truly do seem to be controlled. For example, choosing what to say next or where to go to dinner seems controlled in a way that, e.g., choosing to see when your eyes are open doesn't. Of course, one can choose to open one's eyes or not but once those eyes are open (and one's not blind, and there is some ambient light, and they aren't wearing a blindfold, etc. etc.) one more or less just automatically sees whatever happens to be in front of one's eyes. The same seems to hold for all other modalities too (and language). Affect also seems to be more of the automatic variety. People can be primed for pro-social behavior by smelling freshly baked cookies (Isen and Levin 1972) not because they reason their way

there, but because of the automatic connections between certain stimuli and our affective processes and the automatic connections between those processes and other cognitive processes.

But there are other types of control that aren't comfortably seen as top-down. There seem to be intermediate cases where the locus of control is situated *within* the person, but not at the 'person-level.' Some mental activities are automatically activated (i.e., activated from the bottom-up) but can then be modulated endogenously. One might find oneself tapping one's foot to the rhythm in media res and then decide to (e.g.,) stop tapping or tap faster (the tapping having been started automatically and subconsciously though later controlled in a top-down fashion). Likewise, there is some evidence that we have a visual-postural 'module', one that causes us to automatically orient our body based on our perception of the position of others in our environment (Maisson and Dufosse 1988). If someone next to you is slouching you are more apt to end up automatically slouching over too.[2] Of course, you can also readjust your posture if you care to, so you can exert some control over this process (which is some reason to think that the system isn't classically modular), but the process is set off automatically. These cases identify a different form of control—they are cases where we aren't in control of the activation of a certain process but aren't closed off from some form of steering the process.[3] These processes aren't wholly ballistic and can be partly controlled.

The kind of control I am interested in is a different type of partial control. I am going to meditate on a psychological process, belief fixation, which at first looks like it is of

[2] Likewise, if you are sitting and staring at a wall which is ever-so-gently titling on its axis you will unconsciously align yourself to the tilt of the wall (Fodor personal communication).

[3] For a contrast compare how these cases differ from ballistic processes where one can often control the onset of the process, but once the process starts one can't manipulate it. For example, you can control whether to open your eyes or not (most of the time), but once your eyes are open forming a percept is a ballistic and uncontrollable event.

intermediate case of a controlled process. I assume that the widespread view of belief is as follows: we form beliefs through a process of partial control. In general, we, at the person level, are not in exactly control over what we believe, but some sub-system of ours does exert influence over what we believe. I will suppose that doxastic voluntarism is not the norm, and that most theorists think that we run some type of (generally unconscious) decision procedure in order to figure out which propositions to believe. Thus, although we can't choose to believe, we do have some power over what we believe (of course this 'we' has to be read sub-personally). Note that this is a quite recessed sense of controlled; in a similar sense we have some sub-personal control over what we see, because our unconscious visual processing runs certain computations over the given visual inputs. Just as a (e.g.,) visual module will run certain computations in order to generate a percept, so will out decision procedures run certain computations in order to figure out what to assent to. In what follows, I will argue that, when it comes to first fixating a belief, we don't even have this recessed sense of control over what we believe. The view I espouse is one where original belief fixation is a brute process, one that isn't fruitfully seen as computational; rather I will argue that belief fixation is purely reflexive, not just in it being ballistic, automatic, and mandatory, but also in the Behavioristic sense of it being a non-computational process. My position is that we don't have any control over what we believe once a proposition has been entertained; that is, every proposition that we entertain, we reflexively believe.

Such a view fits well with many other strands emerging from contemporary inquiries into the mind. Seeing people's behavior as deeply impacted by unconscious biases and situational variables has lead to a view of the human condition as much less controlled than even the cynical amongst us might have pretheoretically supposed. But even as the era of the

'Unbearable Automaticity of Being' (Bargh and Chartrand 1999) has shuttled in, our view of our abilities to process the stimuli we encounter (even if such processing is unconscious and automatic) has remained untouched. What follows can be seen as an addition to this more austere picture of the human condition—consider what follows a statement of the Unbearable Automaticity of Believing.

****

I view the picture of belief and belief acquisition that emerges as an amalgamation of many other theorists' blood, sweat, and data. In ending this introduction, I'd like to take care of some intellectual debts I have to pay. My most striking debt is owed to the work of Dan Gilbert (and colleagues) who revived similar view of belief fixation to the one that I'll be advocating here in the late 80's and early 90's. As Gilbert and co. point out, the view that they revive is one that has been around at least since Spinoza. However, as others have been so kind to show me (and without all of the unbelievably generous input of others, there would be no work here to speak of) some of the stoics also seemed to hold a similar view of belief fixation (Long and Sedley 1987).[4]

So, why have I spent my time mulling over a view that has already been held in the literature? Because the view that Gilbert put forth has been both grossly underspecified and underappreciated. In what follows, I don't take on all possible aspects and implications of the view, nor do I straighten out all the details. Rather I deal with what I take the main problems of the view to be. In doing so, I (oddly) find myself following the lead of someone like John McDowell. McDowell sees himself as entering philosophical conversations with the dead only when he sees some problems that could use some housekeeping (McDowell 2009). Like

[4] Thanks to Jesse Prinz for pointing this out to me.

McDowell, I see the insights that I have as a combination of repackaging other's work and fixing some problems and ambiguities in views that I regard favorably. When I stumbled upon Gilbert's work, I was immediately struck by his insights, but also irked by the outstanding issues left unresolved, like what was the nature of belief such that beliefs can be acquired automatically? What was the nature of the introspection such that we couldn't tell that we acquire beliefs automatically? What are the consequences of such a view of belief fixation for other cognitive processes and philosophical endeavors; in particular what is the relation between belief fixation and rationality? Moreover, and of significant importance, Gilbert never distinguished between two competing hypotheses of his data. He never gave arguments for why we should view belief fixation as a reflexive, architectural phenomenon as opposed to a merely heuristic process. Part of my project is to convince you that belief fixation is reflexive and not merely that we have a default heuristic to believe whatever we think. My project will situate belief fixation as a reflexive, architectural phenomenon process that illuminates mysteries in other areas of cognitive science. I hope that this dissertation isn't doesn't just contain arguments for a tendentious view of belief fixation, but rather situates the process of belief fixation within a galaxy of other cognitive phenomena. I hope that after we see how belief fixation works we can see how it solves other outstanding issues in cognitive science and philosophy.

Lastly, the reason I found the Spinozan view appealing is that it comported so well with other pictures of the mind I had been collecting from various sources. I am unbelievably lucky to have had such influences, especially because some of these influences comprise my dissertation committee and have been patient and kind enough to endure me for all these years. I approached the topic of belief fixation from a background of assuming that even

though our experiences can often be crystal clear to us, the workings of our minds are much more opaque than they at first seem. In particular, Fred Dretske' recent work on our lack of introspective access to our mental processes convinced me that the Spinozan view was workable and quite possibly true. Dretske has argued that although we have access to the contents of our thoughts, we don't necessarily have access to the fact that we are thinking (Dretske 2004, unpublished). This insight, combined with a very pregnant footnote from Bill Lycan in his work on tacit belief and Lycan's recent work on cognitive phenomenology (Lycan 1986, forthcoming), made me realize that perhaps we don't have first-person access to our propositional attitudes in general. Both Dretske and Lycan's work convinced me that although we have access to the contents of thought, we don't have access to what types of thought we are having; in other words, we don't necessarily know which propositional attitudes we happen to hold. To use a popular metaphor, we may have a 'belief box' but if so, we don't know what we've put in the box and what's been left out.

Of course, if I think we don't have introspective, direct, 'privileged,' first-person access to our propositional attitudes, I don't think we have access to the types of mental processes we use and the workings of those processes. This insight can be gleaned from many different fields. In particular, I first caught it after spending time with the concept of modularity and in discussions with Jerry Fodor (who I owe a great deal of thanks to for being kind enough to argue with me for so long). I take it that this insight is no longer tendentious, as I find it at work in the research across the cognitive sciences.

My last intellectual debts that I think need to be paid up front are to Joshua Knobe and Jerry Fodor. Both have set an example of how first-rate cutting edge cognitive science and philosophy should be done. They have taught me how to think about cognition, how to

write clear, concise pieces of cognitive science (though of course I often fail to reach the bar they set), and above all, how to argue for radical views in a controlled, sensible way. I can only hope that this document mirrors some small aspects of their work (and if I can't attain their profundity of thought, perhaps I can at least simulate their conciseness). I owe them (and others) more than I could possibly explain, even in a dissertation-length document.

Well then, enough ground clearing. I don't suppose that when we are done the pasture will look cleaner, but hopefully it will at least look different. Or, at the very least, hopefully we'll look past an old pasture and meditate upon a new one, if for only a few fleeting moments. I thank you for joining me on this endeavor. I'll try to keep it as quick and painless as I can.

****

Now for some notes on the structure of this document. This manuscript has 5 chapters. You are currently in the first chapter, which is about to end, so I don't think I need to summarize it further. The second chapter mainly consists of arguments against a ubiquitous and intuitively compelling theory of belief fixation, the Cartesian Theory, and proposes a competing view, the Spinozan Theory. The main goal of the second chapter is to convince the reader that there is something deeply wrong with the Cartesian theory. An ancillary goal of the second chapter, and main goal of the third chapter, is to argue for the plausibility of the Spinozan theory. Specifically, the third chapter is one elaborate abductive argument in favor of the Spinozan theory. The fourth chapter contains objections and replies, and the final chapter contains a sundry array of speculations, observations, and short jokes. It also has a conclusion, but I don't want to say too much about that and ruin it—who would want do such a thing to their readers? Not me.

Before we go any further let me give an apology and a bit more of a road map. This dissertation contains far fewer jokes than I would have liked. For some reason the idea of 'professionalism' has washed all over me and made me painstakingly take out all the most hilarious (or, for people with a different sensibility, 'inappropriate', humor). For this I apologize. I also apologize for the flow of the essay. Chapter 2 is very dense and somewhat boring. There is no easy way around this, for I have to get a bunch of data on the table before the fun can begin. This is why I originally peppered chapter 2 with jokes, but I guess this is a dissertation and I have to pass in order to have the opportunity to eat at the high table and, since I know that the heartiest laugh I can bring you is the thought of me at high table, most of the jokes have been excised. After the second chapter the pace will pick up considerably. I think this gets more interesting and creative as we move along, but then again, my drinking also picked up considerably as the essay continued. Hopefully yours will too

Chapter 2: Two Theories of Belief Fixation

2.1 Impotent Warnings, Disappointed Partners, and Poorly Placed Rocks

Imagine I just heard word of some very fortuitous circumstances: I received a call that my sister's friend is moving out of her rent-controlled New York City apartment. Knowing that this friend thinks well of me, I'm confident enough to believe that if I ask her if I can take over the lease, there is a good chance that she'll agree. However, I also know that she has closer friends than me, who may have the inside track to the apartment.

Knowing all this I call my partner with both excitement and trepidation. I want to let her know about the circumstance, but I do not her to overreact and I fear that by telling her this news she may get her hopes up. Consequently, I begin our conversation by telling her "I've some interesting news, but before I tell you, I do not want you to get too excited." Before telling her the news of the open apartment I convey the probability that what I'm about to tell her probably will not come to fruition. Then after telling her about the apartment I reiterate how it's a long shot that we'll get the apartment. Alas, my warnings do not dissuade her from becoming inordinately excited. After her excitement, she can tell me all the ways in which she sees the apartment going to someone else, but even as she's going through the rational motions, her excitement is never really tempered.

This situation is not an uncommon one. When we give exhortations to others to not get excited at exciting news, these exhortations typically fall on deaf ears even when the reasons for withholding one's excitement are sound. Likewise, if you preface a story that is

apt to make your interlocutor angry with a heartfelt, “Please do not get angry”, this preface is generally as impotent as the request to not get excited at exciting news. Imagine that I just borrowed your car and a sudden and unforeseen hailstorm appeared while I was driving and dropped pieces of hail so thick that, although I took cover as quickly as I could, the hail peppered your car with silver dollar sized dents all over the hood, roof, and trunk. In this situation I may request that you not get angry as shorthand for requesting that you not get angry with me and you may rationally agree that you shouldn’t get angry while still getting upset. Even if I owned a body shop and could fix your car for free, you may still get quite angry just by merely hearing the bad news.

In everyday life we are often faced with situations where we request that someone not to feel an emotion. Sometimes we do this because the person is about to encounter a situation that we know is apt to bring that emotion about, and yet we know the emotion isn’t warranted by the situation. More often than not, our requests are not carried out, even if we have sympathetic listeners.

The situation seems importantly different when we switch topics from exhorting someone not to feel a conative state to exhorting someone not to feel a cognitive state. If I first ask you to please not believe what I’m about to say, it seems intuitively plausible that you will often have no problem not believing what I’m about to say. In contrast, suppose that you have just stubbed your toe on a rock. If you are like most people, you will, at least momentarily, be angry *at the rock*. Even though you might know that the rock is not an appropriate subject of your reactive attitude, you can’t help but be angry at it.

Cases like the one just described are common: we frequently feel emotions that are, even by our own lights, rationally groundless. But we tend to assume that this is not equally

true of our beliefs. If I ask you to please not believe what I'm about to say (because, e.g., I'm merely parroting someone else's falsehood), it seems plausible that you will be able to not believe what I'm about to say. If I tell you that I'm about to read a list of sentences all of which are false and then I read the sentences, it seems plausible that you would not automatically believe these sentences in the way that you may, for example, automatically get excited when hearing of a rare and tantalizing opportunity.

However, in what follows I will argue that this plausible assumption is false: just as we get angry with the rock while knowing full well that it's not an appropriate object of our anger, so too we believe what people say even when we know that what they are saying is false.[5] That is, just as emotions are insensitive to our background beliefs, so too is belief formation initially insensitive to our background beliefs.[6] More specifically, I will argue for the claim that, whenever we entertain a proposition, we automatically believe that proposition. The plausible idea that we can entertain a proposition while withholding assent from it is a myth; it is an idea from which we should withhold our assent.

2.2 Belief Fixation and Rationality

The idea that we can contemplate a proposition without believing it has been accepted in philosophy since at least the time of the ancients and remains widespread in contemporary debates concerning everything from cognitive architecture to epistemology. To take one representative example, Jerry Fodor says,

> To a first approximation, we can assume that the mechanisms that affect [the fixation of perceptual belief] work like this: they [central systems] look simultaneously at the

[5] Or a different formulation for those who think that you can't believe that p and know that not-p: we will believe someone's testimony even while knowing that the testifier claims to be lying.

[6] It's plausible that the process of belief formation is even more encapsulated that elicitation of emotions. I will argue that belief formation is completely informationally encapsulated, so much so that it can be fruitfully seen as completely reflexive.

> 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 (Fodor 1983, p.102).

Note that this story assumes that our central systems examine how different entertained propositions are analyzed in light of our background beliefs. Fodor assumes that background beliefs interact with propositions we entertain because he thinks that belief fixation is a rational, conservative, gradual, slow process that takes into account all the relevant data in one's information store before assenting to any proposition.[7] Here (for once) Fodor's view is quite indicative of the field at large. Belief fixation is hypothesized to be a slow, conservative process to, in part, allow for the idea that we have the ability to contemplate the truth of a proposition before assenting to that proposition. This intuitive view is at odds with a theory where any proposition that is entertained is just automatically and reflexively believed. So, if it were true that belief fixation was reflexive (such that every thought that was contemplated was believed) and interacted with no background information, then it would be a very interesting and surprising fact about the mind.

The consequences of such a radical departure from the standard view would extend far beyond the topic of belief fixation. The ability to withhold assent from propositions that we entertain is a crucial part of our picture of impartial doxastic deliberation (i.e., the ability to impartially consider propositions while suspending judgment). We take ourselves to be able to consider propositions while remaining neutral as to their truth. Furthermore, impartial doxastic deliberation is integral to our conception of what it is to be a person because we take people to be paradigmatically rational creatures, and impartial doxastic deliberation appears to be a necessary condition on rationality. If we found creatures that couldn't help but believe

[7] Hence Fodor writes things like "the fixation of perceptual belief is the evaluation of such hypotheses in light of the *totality of background theory*" (italics added, Fodor 1990, p.248).

any idea that they entertained, we would be inclined to regard them as massively irrational. Sadly, we seem to be such creatures.

To pump your intuitions a bit, 'imagine' that we found out that all of Queens was inhabited by aliens. Now suppose that one day you find yourself wandering around Queens looking for the closest subway. You pass one of these aliens and ask which direction is the closest subway. The alien then tells you to walk two blocks east. Being a bit unsure of these aliens (after all, they live in Queens) you ask the alien to give some justification for 'his' belief. Suppose 'he' responds by saying, "I don't have any reason to believe it, I just do. I just remember once thinking that the subway was east of here so I believe it. But I also believe that the subway is west of here. Anything that I happen to think, I end up believing." I don't think that one would find such a creature to be the paradigm of rationality.

If you share the intuition that such a creature shouldn't be considered rational, then you should agree that there is a connection between how one forms beliefs and how rational one is. Later in this essay, I will contend that our view of ourselves as rational creatures is imperiled by the way in which we process information. A critique of rationality stemming from our inability to rationally deliberate differs from the contemporary "rationality wars" criticisms (Samuels et al. 2002). Recent decades have brought heated debates over how rational people are, but these debates cluster over whether people tend to answer some particular problem correctly. One needn't look hard to find claims that people are irrational because they fall prey to cognitive illusions, use fast and frugal heuristics, let emotions dictate their moral reasoning, etc. Throughout these debates, a cornerstone of our rationality has remained beyond critique: our ability to contemplate propositions without believing them. This ability has received scant attention and has endured few serious critiques. Yet

when one looks closely at our actual doxastic capacities, the picture that arises is surprising and quite epistemologically troubling.

If the theory I propose is correct, then we will have to reconsider the nature of doxastic deliberation and whether we are able to engage in it. This is because if the proposed theory is correct, then impartial doxastic deliberation is impossible. Consequently, the theory of belief fixation I propose is somewhat radical and unintuitive. My goal is not to establish the truth of the theory beyond a doubt, rather my aim is more modest: to convince you that it is a plausible model of our cognitive architecture that demands further investigation. And even more modest goal is to merely get you to understand my model of belief fixation, for so long as you do, you will believe it.

2.3 Beliefs, Credences, and Functionalism

But before we get there, let's first be clear about what notion of belief we'll be working with and what metaphysical pretensions (or lack thereof) I have. The notion of belief that is operative throughout this paper will be the quotidian notion of belief that is operative in the cognitive sciences, with belief understood as a relational, gradable, functional state. This notion of belief, being gradable, allows that one can believe things to stronger or weaker degrees. For current purposes, belief will not be understood as merely a binary relation where one either does or does not believe that P.[8] Rather, belief will be understood similarly to the way one understands credences.[9]

[8] I say 'merely' because the gradable notion still allows for some binary notion of belief.

[9] Thus, one can interpret my theory as stating that whenever you entertain a proposition, you raise your credence in that proposition. How high is credence raised? Is it to a high degree or just to a non-zero degree? To a first approximation, the credence is raised to a level that would generally produce behavior. Presumably a belief with a credence of .0001 won't produce much if any behavior; on the other hand, a belief needn't have a credence of .9 in order for the belief to have behavioral consequences. I take it as an open empirical question how high one's credences have to be for a belief to regularly eventuate in behavior. The operative claim in the text is that entertaining causes one's credences to go at least that high.

Now for a note on what I'm decidedly not doing: I'm not arguing for any metaphysically necessary conclusion. I will be happy to 'settle' for nomological necessity. Figuring out the generalizations about minds here on earth is a hard enough problem that I don't need to worry about what minds must be like on Pluto. I'm more interested in partaking in some speculative psychological theorizing than I am at analyzing what our concept of belief is. Furthermore, I don't especially care what one calls beliefs. I'll argue that the things I will call 'beliefs' do all the same work as the things the folk call 'beliefs.' What's important is not what we name the thing, but rather that we accept that the type of mental state I'm discussing is a true psychological kind and helps to explain a significant amount of human behavior. I see myself as not eliminating beliefs, but instead just clarifying and filling out some of the properties of belief. I take it that all, god fearing, functionalists thought that sooner or later we'd fill in the dummy properties of belief and one can see the account I offer as an example of that filling in.[10]

Though I'd like to stay neutral about some more exotic topics about belief, I will need to specify a bit about beliefs so that my view. To a first approximation, beliefs are functionally individuated cognitive states that are truth-evaluable and are causally efficacious. They have some canonical causal powers: beliefs can interact with other mental states and cause both the generation of still further beliefs and behavior. The main sense of 'belief' at use in the cognitive science literature is of a mental state that interacts with (at the very least) one's desires to eventuate in behavior. One needn't be a behaviorist to see that the connection to behavior is integral for something being a belief. If we claim that person X has a belief but that the belief would not under any circumstance eventuate in behavior, I think

[10] For arguments that these states are actually beliefs and not some other cognitive state, see section 4.5.

we'd be very skeptical of calling that thing a belief. Regardless of what we might do in this situation it's pretty clear that such a state could never count as a belief in modern cognitive science. In what follows, I will argue that the states I discuss have all the aforementioned, necessary properties of belief.[11]

In the rest of this chapter, I compare two theories of belief fixation. The goal of this chapter is to convince you that there is something deeply flawed with the intuitive, widespread view of belief fixation. After persuading you thusly, I will then give an abductive argument in favor of a competing view of belief fixation, then deal with objections and critiques. But for now, let's peruse a ubiquitous and influential theory of cognitive architecture: the Cartesian theory of belief fixation.

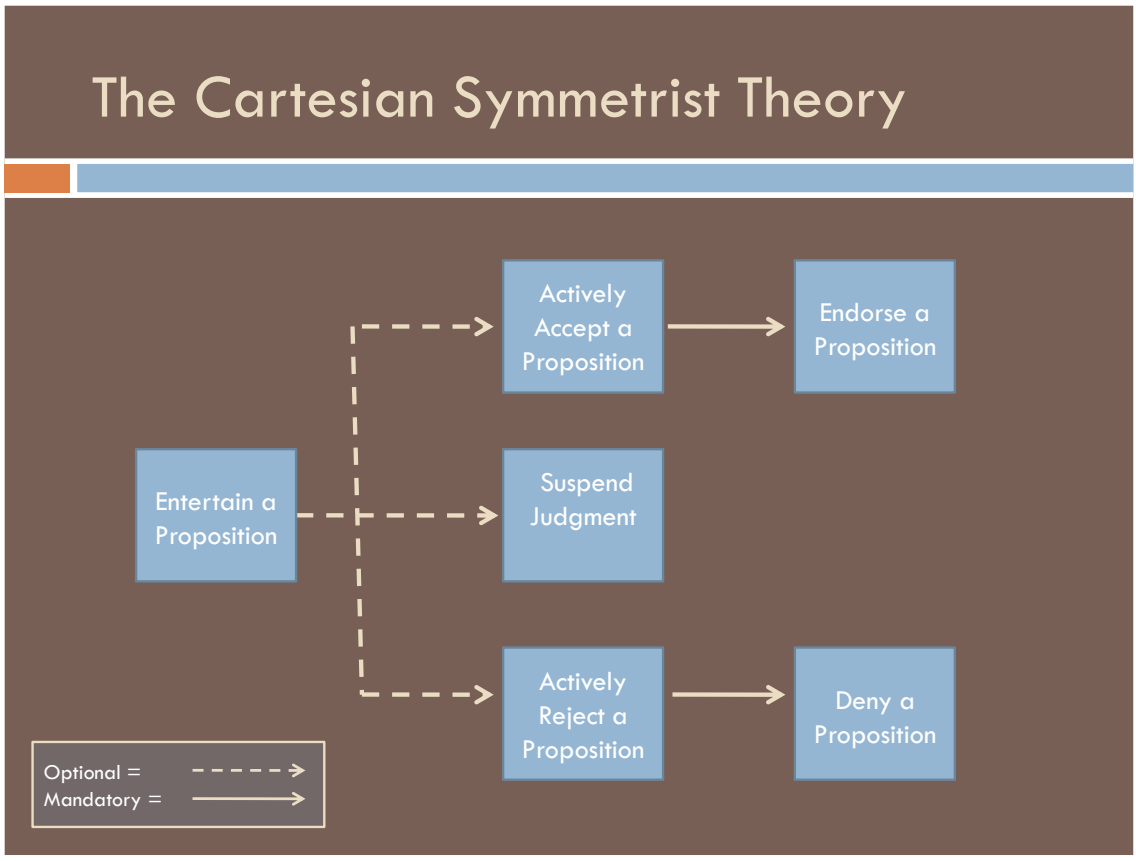
2.4 The Cartesian and Spinozan Theories of Belief Fixation

The methodical withholding of assent is part of a venerable epistemological tradition: if surety is what one desires, then one should be skeptical of what one thinks, waiting for the ideas that pass through one's mind to be 'clear and distinct,' or at least well justified. Surety was Descartes stated goal in the *Meditations* (1641/1988). But it's worth asking: when Descartes was sitting beside the fire contemplating which propositions to believe in, what was he actually trying to do? He was attempting to first *entertain* an idea, then *contemplate*

[11] They also seem to have other properties that other theorists have offered as necessary conditions on belief. For example Fodor, states a few conditions that propositional attitudes must meet: they must be analyzed as relations (thus keeping appearances of the English sentences that they parallel, which look to be relations), account for opacity, have logical form, and mesh with empirical accounts of thought processes (Fodor 1981a). The beliefs posited by the Spinozan hit all of these conditions: they instantiate a relation between a person and internal formulae (the latter which itself bears a relation to certain propositional contents). By exploiting the internal formulae story they can also account for opacity and logical form. As for the last condition, the empirical plausibility, the majority of this essay will be spent arguing that the Spinozan beliefs, but not the Cartesian ones, satisfy this condition. Although I've identified a slew of putative necessary conditions of belief, it may be the case that there are some further conditions that need to be met in order for something to suffice as a belief. I cannot list these further conditions, but I bet no one else can either (though I am open to suggestions). It is well known that one can specify necessary conditions on the cheap, but specifying a sufficient condition is damn near impossible and this point is a general one that holds outside discussions of belief (Fodor 1981b).

its truth, and finally *decide* what to assent to and what to *withhold* judgment from. Descartes' venture presupposed a serial model of belief fixation, according to which one first entertains a proposition, then subsequently believes, rejects, or withholds assent from the proposition.[12] This type of serial model presupposes that a) the faculty of entertaining a proposition is a separate faculty from the faculty of believing a proposition and that b) the workings of the former faculty are prior to the workings of the latter (see Figure 1). These assumptions are at the heart of the serial model of belief fixation, which I will term 'the Cartesian theory of belief fixation.'

[12] Although this scenario admittedly paints Descartes with a broad brush, some relevant literature has interpreted Descartes as attempting the project I sketched out (e.g., Gilbert 1991, Huebner 2009). Nevertheless, there are some reasons to believe that Descartes actually wasn't a Cartesian in this sense. Historians like Alan Nelson (personal communication) interpret Descartes' epistemic methodology as such: assume Descartes wants to assess the truth of the proposition that Santa Claus exists. Call this proposition S. Descartes' first step in assessing S is to token the thought WITHHOLD ASSENT FROM S (actually Nelson's take on this seems to be that the first step is to token the thought: THINK WITHHOLD ASSENT FROM S; I'll ignore this element, which strikes me as regress prone.) The next step is to think of situations which would entail the falsity of S—for example, imagining an empty North Pole. The reason we think of an empty North Pole as opposed to thinking NOT S is that Nelson's Descartes doesn't believe one can just think of negation as such (nor presumably does Nelson's Descartes believe we can think *with* negation as such). Nelson's Descartes holds a variation on the view that I'm promoting; he holds that people believe everything they think because they do not have the ability to withhold assent. What people can instead do is constantly have a belief swamped by a contrary belief. In essence, this reading of Descartes interprets the withholding of assent as a type of thought suppression: your belief that S is weak if it immediately leads to a different belief and it is super-weak if it leads to a different belief that would entail the falsity of S. A strong belief is a belief that doesn't automatically lead to a second belief, which destroys our consciousness of the first belief. So, perhaps Descartes wasn't a Cartesian in the sense expressed in the main text (of course, as Dan Garber fervently claimed to me, perhaps he was). That doesn't really matter because an overwhelming majority of contemporary philosophers and cognitive scientists are. If one would like they can substitute Pollock (1986) or Fodor (or anyone else who has the modular/central systems distinction) in for Descartes (see Fodor 1975, 1983, 1998).



(Figure 1.) (NB: The dotted lines represent optional links, and solid lines necessary links)

What I am here calling the "Cartesian theory of belief fixation" consists of the following claims:

1) People have the ability to contemplate propositions that arise in the mind, whether through perception or imagination, before believing those propositions.
2) Accepting and rejecting a proposition exploit the same mental processes, and consequently, should be affected by performance constraints in similar ways.[13] I will sometimes refer to the Cartesian position as a 'symmetrist' position because it treats accepting and rejecting symmetrically.

[13] I'll use the phrases 'accepting a proposition' and 'believing a proposition' interchangeably; likewise for 'rejecting a proposition' and 'disbelieving a proposition' (though I find 'disbelieving' to be both awkward and ambiguous, so my use of it will be quite sparse).

3) Forming a belief is an active endeavor. Since accepting a proposition and rejecting a proposition are underwritten by the same mental processes, rejecting a proposition is also an active endeavor.[14]
4) One can hold a belief and then later decide that the particular belief is false and consequently stop believing for broadly rational reasons. Thus, just as a person can acquire beliefs for rational reasons, a person can lose beliefs for rational reasons.

The Cartesian theory is intuitive, widely accepted, and rarely, if ever, argued for. It is assumed throughout seemingly every area in cognitive science and is shared by philosophers who have radically different views of belief. For example, one can see the Cartesian view assumed in Interpretationist views of belief (e.g., Davidson 2001; Dennett 1987) or Realist views of belief (where realist for present purposes need only mean that the facts about belief outstrip facts about idealized interpretation e.g., Fodor 1975, 1983, 1998);[15] it can be found in theories put forth by social psychologists (e.g., Festinger 1957; Milgram 1974; Cooper 2007) and cognitive psychologists (e.g., Pylyshyn 1989a; almost every author in Ford and Pylyshyn 1996).[16] However, there is mounting evidence that the Cartesian theory is more

[14] Suspending one's judgment can be either active (as when one decides that there is not enough information to decide one way or the other) or passive (as when one's head becomes momentarily attached to a fast moving brick, thus making the decision process moot). On the Spinozan theory, even a fast moving brick couldn't derail one's passive assent.

[15] Of course, realism about belief generally involves quite a bit more commitment. For example, Fodor, an arch-realist, will also aver that beliefs are concrete mental particulars with robust causal powers. Contrarily, an interpretationist like Dennett will state that although beliefs have causal powers (see Dennett 1991), they are not concrete mental particulars (I find this position to be quite unstable).

[16] As aforementioned, Fodor explicitly presupposes the view as part of the distinction between central cognition and modular input systems. Pylyshyn (and most of the authors in his co-edited volume) also assumes a similar dichotomy to Fodor's (often with talk of 'modules' being exchanged for talk of 'cognitive impenetrability,' which amounts to about the same thing). Dennett and Davidson implicitly assume the view as part of their principle of charity (or intentional stance): if beliefs are posits based on what is rational to ascribe someone,

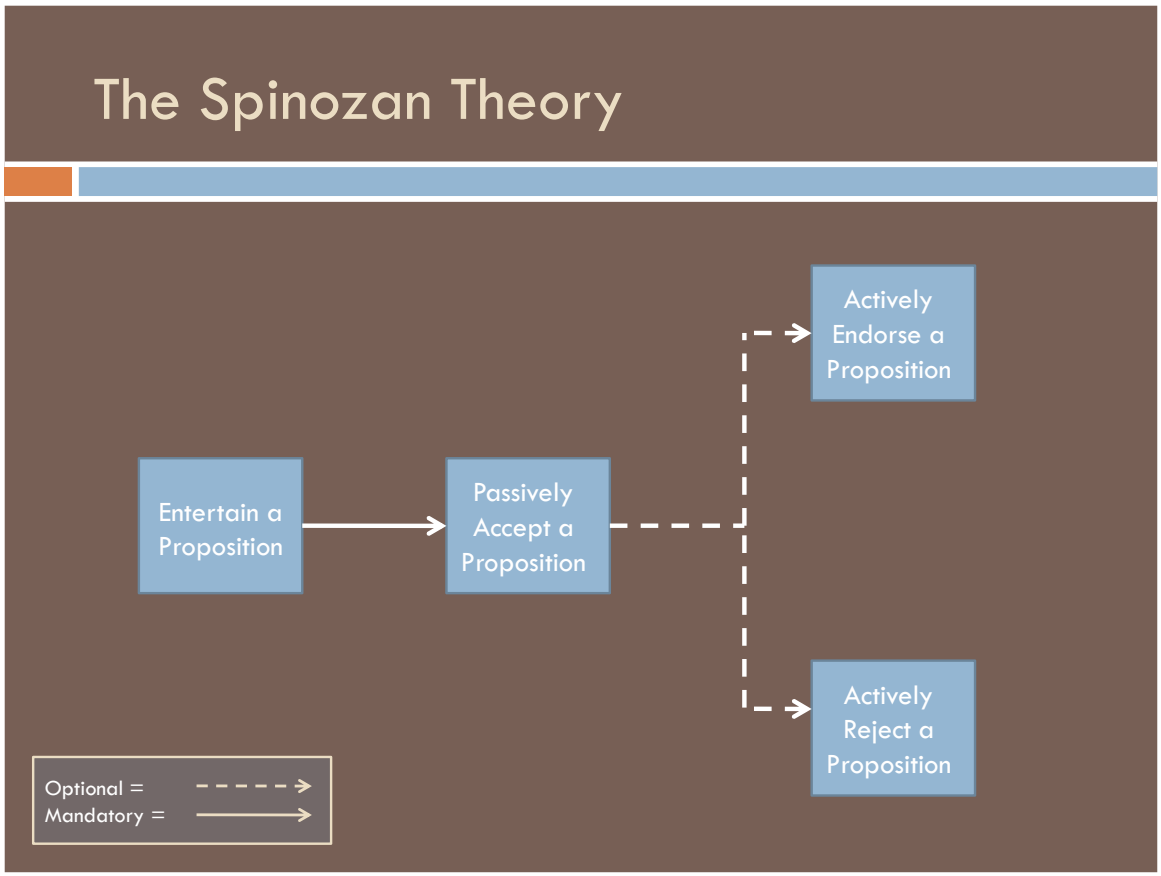
venerable myth than hard fact. Consequently, we can be thankful that the Cartesian view isn't the only available theory of belief formation. Spinoza (1677/1991) had a competing view of belief formation, one according to which contemplating a proposition's truth coincided with assenting to a proposition. In lieu of the Cartesian view, I propose a version of a Spinozan theory of belief fixation, one in which tokening[17] an idea is sufficient for believing that idea.[18] For the Spinozan theory, one automatically and passively accepts whatever ideas one tokens, and only after the initial acceptance can one effortfully reject one's newly acquired belief (see Figure 2).

---

then we only posit beliefs that would be, on the whole, rational for the believer to hold. Which means we shouldn't ascribe beliefs to people based on the fact that they merely contemplated a proposition; rather, we should ascribe beliefs to people assuming that they consider certain propositions and only believe what they have some decent (in a subjective sense) evidence for. Though Festinger and Cooper both ascribe to different varieties of dissonance theory (Festinger's being the basic version as opposed to Cooper's "New Look" version), both variations assume that one forms beliefs based on one's previous beliefs and emotional commitments, thus implicitly ascribing to a view that propositions can be entertained in light of one's previous beliefs and commitments (even if the decision procedure brought to bear on such propositions is less than rationally satisfying, it is a decision procedure nonetheless). Milgram accepts the Cartesian view for queerer reasons. Milgram thinks that the default state is for people to be able to entertain propositions as long as overriding situational constraints (like social pressure) don't intervene. Thus, Milgram probably would have found the 'gullibility heuristic' story quite tempting (see section 4.2).

[17] I use 'tokening' because it strikes me as the most neutral and general verb for covering the category of heterogeneous mental acts addressed by my theory. These acts include understanding, entertaining, contemplating, and related activities. If you are having trouble envisioning the thesis assume that there is a language of thought (LOT). My thesis is that every time a truth-apt sentence is tokened in one's LOT, one believes that sentence.

[18] Having no metaphysical axes to grind (here at least), I don't particularly care whether we believe propositions or whether we believe ideas. I will thus use these descriptions interchangeably. The difference between the two does not affect my main points, but if you prefer you can substitute each for the other throughout. For what it's worth, it sounds most natural to me to say that beliefs are propositional attitudes that instantiate a particular relation (some of the properties of which will be discussed later) between a thinker and a set of mental representations (which presumably are, but needn't necessarily be, concepts). It thus makes sense to my ears to say that we believe propositions, but beliefs are 'made out of' (read: the instantiation of a certain relation to) mental representations (a structured set of concepts) which themselves express propositions.



(Figure 2) (NB: The dotted lines represent optional links, and solid lines necessary links)

What I am here calling the "Spinozan theory of belief fixation" consists of the following claims:

1) People do not have the ability to contemplate propositions that arise in the mind, whether through perception or imagination, before believing them. That is, because of our mental architecture, it is impossible for one to withhold assent from propositions that one tokens.[19] Thus, one can never suspend judgment.

[19] The impossibility claim is there to rule out that one has a heuristic that makes people tend to believe what they perceive (for more elaboration on this point and a response see the 'Gullibility Heuristic' entry in section 4.2). N.b., the previous statement is not meant to imply that people actually perceive propositions. Rather, a phrase like 'believing what you perceive' is shorthand for 'believing what normally comes to mind when you perceive X.' The idea behind this is quite tame: many perceptual situations lead to the corresponding automatic tokening of thoughts. Maybe one can perceive propositions (though saying that sounds odd to me), what is important here is just to note that my account needn't take a stance on this topic.

2) Accepting a proposition is accomplished by a different system than rejecting a proposition. Because different systems are at play, the processes of accepting and rejecting should be affected by performance constraints in different ways. I will sometimes refer to the Spinozan position as an 'asymmetrist' position, because it treats accepting and rejecting asymmetrically.
3) Forming a belief is a passive endeavor. However, rejecting a proposition is an active and effortful mental action, which can only happen after a belief has been acquired. Consequently, one can effortlessly form new beliefs while being mentally taxed, but rejecting an already held belief will become more difficult the more mentally taxed one is. For the Spinozan, every proposition that is entertained is necessarily accepted, but every proposition that is accepted is not necessarily endorsed.[20]
4) Losing a belief is never a purely rational process. That is, even if you clearly see the falsity of your belief that P, you still can't just stop believing that P. This point is somewhat tangential to the core of the overall theory. According to the current taxonomy, one could not be a Spinozan without endorsing the first three properties enumerated, but the rest are optional.[21]

My version of the Spinozan theory takes on an extra commitment on an issue about which the Cartesian theory is agnostic: the relation between rejection and negation. Because

[20] For current purposes, endorsing a proposition is something that happens at the person level. One consciously chooses what to endorse, whereas accepting needn't be conscious nor volitional. In the Spinozan ontology, denying is the negative complement to endorsing (also a person level phenomenon), whereas rejecting complements accepting (and both are sub-personal phenomena).

[21] Some suggestive evidence for this suggestion comes from a recent study that shows that priming effects lasted on subjects 17 years after the original prime (Mitchell 2006). However, since the evidence in favor of (or against) such a view is so incredibly scant, arguing for this property from empirical data is a very difficult endeavor. In section 5.2.2, I will offer some suggestions in favor of holding the view that one can never lose beliefs for rational reasons.

the Spinozan theory dictates that accepting and rejecting are subserved by different mental processes, it's natural for such a theory to give some analysis of what rejecting is. As opposed to analyzing rejection in terms of negation, I follow Price (1990) in inverting the direction of analysis and instead analyzing negation in terms of rejection. The following property (and consequences thereof) is thus not one that every Spinozan theory must adhere to; rather, it is idiosyncratic to my own version and can be evaluated separately from the first three properties.

5) To negate a thought is to, in part, reject it.

Now for a few non-obvious consequences of the Spinozan view. The Spinozan sees acceptance and rejection as different propositional attitudes. However, the logical relations between these attitudes can differ based on one's tastes. For example, a Spinozan who denied property 5) could hold that accepting not-p does not entail rejecting p, though a Spinozan of my variety has to allow the entailment (the reason being that a Spinozan of my variety always buys that negations are a subset of rejections, so any time something is negated, something is rejected). However, no Spinozan can allow that one can reject p without also accepting p (because anytime one rejects p one first thinks p and thinking p suffices for believing p).[22] Consequently, any Spinozan will predict that people believe many contradictions.[23]

[22] This statement can't be counter-exampled by the case where one thinks, REJECT, for in that case a) one isn't actually rejecting anything and b) it's a non-propositional thought anyway, thus outside the scope of the current discussion.

[23] Of course, this does not entail that people will *assert* contradictions. What one asserts is tied to what one endorses and endorsements are a species of judgment, not belief (for more on the relations between endorsing/denying and believing/rejecting, see the end of section 4.1). One may object that it is incoherent to attribute contradictory beliefs to people. After all, it is commonplace to think that in cases of intentional action people are disposed to act in ways that would bring about their desires if their beliefs were true. But if we had a person with inconsistent beliefs how could she possibly act in a way to fulfill her desires? As Egan (2007) writes, "Which of the actions available to me are the ones that would (tend to) bring about the satisfaction of my

As per property 3), exercising the faculty of rejection is effortful. However, the Spinozan does not predict that rejection is effortful merely because it is the second step in the system; rather, rejection is effortful because the connection between acceptance and rejection is not mandatory. For current purposes, all mandatory processing connections should be thought of as effortless and all non-mandatory processing connections as effortful. This is because all mandatory connections are automatic, like a reflex.[24]

As per property 5), negating involves rejecting. Since negating involves rejecting, and since, as per property 3), rejecting is effortful, negating is effortful too. Thus, negative sentences/thoughts should be more difficult to process (e.g., take longer and be more error prone) than affirmative sentences/thoughts. Furthermore, because negation involves rejection and because one can only reject complete propositions,[25] the Spinozan theory predicts that

desires if P and not-P"? (p. 5). However, this objection is unpersuasive for a few reasons (one being that people as a matter of fact do often harbor inconsistent beliefs, sometimes the inconsistency is just fragmented amongst different belief systems a la Lewis 1982). Most germane here is to note that even if people do harbor inconsistent beliefs they can still hold these beliefs to different degrees and have one more salient (and thus more active) in certain circumstances and others more salient in other circumstances. For more on the contextual factors needed in actual belief ascription see Egan (2007).

[24] Thus 'effortful' is a term of art and as such will deviate from folk usage. An activity will be deemed effortless if it can proceed at normal capacity even if other effortful cognitive activities are occurring and an activity will be deemed effortless if it cannot proceed at its normal capacity while another effortful process is occurring. For example, seeing is effortless because one can see while (e.g.,) solving algebraic equations. Solving algebraic equations is effortful because one's performance will drop precipitously if one tries to solve algebraic equations while also trying to play chess (which is also effortful because one's chess playing proficiency will also drop when engaged in other effortful activities, like counting backwards from 100). In how I will use the term 'effortful', certain activities can be cognitively effortful yet not *feel* effortful. Planning what one will do with one's evening is effortful in the same sense that planning what one will do with one's life is effortful, but the latter and not the former may feel effortful. My sense is that the feeling of effortfulness is in part based on one's motivation for engaging in the activity: one might enjoy planning one's evening, but it's probably rarer for one to enjoy planning one's life (although I've, ahem, heard of certain experimental philosophers who enjoy doing these things on New Year's Eve).

[25] One can reject the proposition *that bear is made out of ice cream*, but one cannot reject a sub-propositional structure like *bear*. I'm not sure there is any sense to be made of rejecting non-propositional structures.

negations can only be processed after a complete (affirmative) proposition has been formed. As a consequence, negations should be processed last when processing negative clauses.[26]

As a consequence of properties 2) and 5) the Spinozan view not only treats acceptance and rejection asymmetrically, but also treats negation and affirmation asymmetrically. The Cartesian position officially makes no predictions about negations, but it's quite natural for Cartesian to be a symmetrist about negation as well as belief.

The Big Picture: on the Spinozan view, any propositional thought one tokens, one thereby believes. Only after a belief is acquired can decision procedures be brought to bear on the belief. If one tokens a dubious proposition, one can effortfully attend to the proposition and reject it. Further contemplation can toggle the strengths of these beliefs, reducing the strength of the affirmative belief and raising the strength of the negated counterpart.

****

The Cartesian and Spinozan theories make quite different predictions. If the Cartesian view is right then we should be able to dismantle the belief-fixating process after the understanding has happened but before the believing (or disbelieving) has occurred. In such a case the Cartesian view predicts that the system will be agnostic about the truth of the proposition. Consequently, since cognitive load is a disabling performance constraint, the Cartesian theory predicts that deciding about the truth of a proposition should not normally

[26] Importantly, the claims in the text regarding negation do not pertain to syntax; rather, they pertain to understanding negation. Additionally, the claims about negation apply to propositions, not necessarily sentences. So, for example, the theory handles embedded negations like the one in 'John believes that Jesse is not a communist' by stating that the negation is processed after the clause sans negation (i.e. 'Jesse is a communist') is processed, not after the entire sentence sans negation ('John believe that Jesse is a communist') is processed.

occur under cognitive load. Additionally, because the Cartesian theory treats assenting and rejecting identically, it predicts that cognitive load will affect both processes identically.

In contrast, if the Spinozan view is right, then the belief-fixating process can be dismantled by invoking some performance constraints prior to rejecting a proposition, but never before accepting a proposition (assuming the proposition is tokened in the first place). Because the Spinozan theory posits that believing is reflexive, believing should occur even when one is under cognitive load. Since the Spinozan view treats accepting and rejecting differently, with rejection being effortful, it predicts that load should only affect rejecting a proposition, not assenting to it.

We will return to these predictions throughout the paper. For now, let's turn our attention to some evidence that should make us quite wary of the Cartesian view.

2.5 Death by 10,000 Murderous Raindrops

I'm going to use the time-tested 'kill you with 10,000 raindrops' approach. This approach is currently quite popular in cognitive science (see, e.g., Doris 2002), but I think my approach differs slightly from others. The approach is generally used when one piece of evidence alone is not sufficient for deriving the desired conclusion. But I think each piece of evidence I will present is by itself strong enough to argue against the Cartesian view. Each piece of evidence below is evidence that the Cartesian theory cannot account for and the Spinozan theory can. Some are more suggestive than others, but combined they make for a daunting challenge for a Cartesian theory.

2.5.1 Memory Asymmetries between Truths and Falsehoods

The most paradigmatic anti-Cartesian experimental paradigm is one that exploits asymmetries in people's memory of truths and falsehoods. In a typical experiment,

participants are asked to take part in a learning task while being intermittently put under cognitive load. After the learning phase, they are then subsequently tested about what they learned and asymmetries in the participants' performances appear. In one telling experiment participants were asked to learn nonsense word meanings. They watched a computer screen where sentences of the form "An X is a Y" appeared, where the "X" was a nonsense word and the "Y" was a word in English (for example, "A suffa is a cloud", from Gilbert et al. 1990). Right after participants read the sentences the screen flashed either the word 'true' or the word 'false,' indicating whether the previous statement was accurate or not. Participants were told to be on guard for a tone that would occur; the tone would occasionally bellow and when it did the participants were to push a button as soon as possible. The tone task was introduced in order to induce cognitive load. On the critical trials participants read six true and six false claims, four of which (two true, two false) had the tone go off interrupting their reading during the trial. At the end of the trials the sentences were then turned into questions (e.g. "Is a suffa a cloud?") and the participants answered accordingly.

The Cartesian view predicts that the tone task should affect both true and false statements equally since although contemplation has occurred, the participants haven't yet had the time to integrate the information properly because of the cognitive burden brought on by the tone task. The Spinozan view predicts that during interrupted trials participants should mistake false claims as true, but not true claims as false, the reason being that the belief fixating system's processing gets shut down by the cognitive load *after* comprehension but *before* rejection. The Cartesian view predicts incorrectly: the added cognitive load made

participants reliably encode true statements as true, but consistently incorrectly encode false statements as true.[27]

This type of asymmetry can be seen throughout the literature: a person put under cognitive load is apt to remember statements that they are told to be false as true but not statements they are told to be true as false.[28] For another very similar example consider the following experiment. Participants learned about the physical properties and habits of an imaginary animal called a "glark." Participants were seated in front of a computer screen that had a camera mounted on top of it. They were told that the camera was an eye-tracking device, though it wasn't (the camera was just there to sell the cover story). During the learning phase participants were shown affirmative sentences about glarks (e.g. 'glarks have white fur'). The testing phase commenced after the participants had read the twenty sentences. The testing phase consisted of eighty sentences. Each sentence was preceded by a prompt which told the participant to either assess the truth of the sentence or commanded the participant to speed read the sentence. The fake eye-tracking device was there to ensure that subjects would actually read during the speed reading trials (participants were told that the camera would check that they actually did read the sentence).

During the first testing phase participants were given fifteen true sentences (e.g. 'glarks have white fur'), fifteen false sentences (e.g. 'glarks have brown fur'), and ten meaningless sentences (e.g. 'glarks have tired fur,' these were given to keep the participants sharp and only appeared during speed reading trials). The syntax of the true sentences was

[27] Participants answered correctly on the true statements 55% of the time when uninterrupted and 58% of the time when interrupted, but participants answered correctly on the false statements 55% of the time when uninterrupted but only 35% of the time when interrupted (Gilbert et al. ibid).

[28] If you found the first study convincing, feel free to skip the next two studies described and instead go to the end of this subsection (and if you really feel convinced feel free to skip ahead to the concluding chapter).

never identical between the learning and testing phases (ensuring that participants couldn't respond based on mere syntactic recognition). Of the true and false sentences, some were given during the speed reading trials and some during the comprehension trials. The crucial evidence came during the second testing phase. During the second testing phase, participants would assess sentences from the true or false categories that they previously had to speed read, not assess, on the first trial. The Spinozan model, but not the Cartesian model, predicts that participants would be more apt to mistake the false sentences they merely read during the first trial as true during the second phase. The Cartesian model, meanwhile, should predict no difference between true and false sentences.

Again, the Cartesian theory misses the asymmetry between one's encodings of true and false propositions. When participants speed read (in round 1) then assessed (in round 2) true sentences, they were more likely to answer correctly (that is, they were more likely to remember the true sentence as true), whereas when participants speed read false sentences (in round 1) and then assessed them (in round 2), they were significantly more likely to misremember the false sentence as true ones.[29]

I'll mention one other similar experiment just to drive the point home, and then leave the issue be. A study with a very similar moral was run on participants who were presented with statements about a non-sense language that the experimenter pretended was Hopi (Gilbert et al. 1990). Subjects read sentences of the form 'An X is a Y' (e.g. "A dinca is a flame"). After they read the sentences, the participants were told that the sentence they read either expressed a true or a false proposition. Some of these trials included a tone which would come up right after 'true' or 'false' was flashed on the screen. Participants were told to

[29] Gilbert et al. (ibid.).

be on guard for the tone and asked to push a button to make the tone dissipate, thus intermittently invoking cognitive load on the participants (the load again being induced on the trials when the tone arrived). When the subjects were later tested the cognitive load was seen to greatly affect their ability to identify false sentences as false. When under load, participants were twice as likely to identify false sentences as true, than they were to identify true sentences as false. This type of asymmetry is expected on the Spinozan hypothesis, but not on the Cartesian view.

This robust asymmetry helps to confirm the second and third properties of the Spinozan theory. The experiments above display that accepting a proposition (here tested as remembering the proposition as true) comes much easier than rejecting a proposition (here tested as remembering the proposition as false). Accepting is easier because it is a passive process, whereas rejecting is an active one. The added cognitive load helps to shortcut the active rejection, but does not interfere with passive acceptance because the passive process is automatic and load does not affect a reflex. Compare how counting backwards from a hundred by increments of five would affect *seeing* a crossword puzzle vs. *completing* the puzzle. The former will not be affected while the latter will be greatly affected. Rejecting a proposition is more like thinking than seeing, while accepting is more like seeing than thinking.

In sum, the observed asymmetry can be predicted if we assume what the Spinozan view asks us to: that when propositions are initially processed they are encoded as true and can only subsequently be marked as false. Not only does the Cartesian view miss the asymmetry between acceptance and rejection, but it misses that acceptance is automatic. The Cartesian view predicts that load should shut down acceptance because it assumes that

acceptance is active. Contra the Cartesian view, load seems to increase, not decrease, people's disposition to accept propositions.

2.5.2 Passive Belief Formation: Inferences from Automatically Acquired Beliefs

One might think that the aforementioned studies are a bit queer because they all involve nonsense stimuli. Perhaps, one may object, participants are unmotivated to deal with nonsense stimuli in the ways the experimental design demands. Although such an objection strikes me as unmotivated, there are some data that speaks against such an interpretation while also creating trouble for the Cartesian view (and supporting the Spinozan) in other ways. The next two studies I will describe will all contain similar morals to the memory asymmetry studies with two twists: the prompts involve social stimuli (thus increasing ecological validity), blunting the above worry and the studies involve the participants making *inferences* from the passively accepted propositions they've acquired, showing that the experimental data reflect the participants' beliefs and not some other mental states.

2.5.2.1 Passive Belief and Criminal Sentencing

The first example we'll analyze deals with participants who were asked to watch a video screen with two crawling scrolls on it, one on top of the other.[30] The top scroll contained text reports of two unrelated crime incidents. Participants were told that they would read both true and false details about the incidents, true statements appearing in black, false statements appearing in red. The bottom crawl did not contain any text, but instead had digits that slowly moved across the screen. Half the participants were told to ignore these digits (the unburdened participants) whereas half were told to peruse the digit crawl and to push a button anytime the number 5 appeared (the burdened participants).

[30] Gilbert, Tafarodi, and Malone (1993).

At the conclusion of the video, participants were asked to recommend a prison sentence for the offenses, ranging from zero to twenty years and were also asked to assess the criminal's personality (in particular, how much they liked them, how dangerous they were, and how much counseling would help them). The false statements the participants read during the first phase of the experiment either exacerbated or mitigated the severity of the crime. The participants in the burdened condition were significantly more likely to be persuaded by the false information. The participants in the unburdened condition recommended a sentence of six years when the false information was extenuating and seven when it was exacerbating (not a significant difference), whereas their burdened counterparts recommended five years in jail in the extenuating condition and eleven years in jail in the exacerbating (a significant difference). Significant differences were also found across the board when looking at the defendant's likeability, benefit from counseling, and dangerousness. Thus, it appears that the burdened participants believed the lies they read more than the unburdened.

It is worth noting that both groups read the same lies, so if it was just associative connections that were affecting the groups the groups should be affected equally. The only difference between the two groups is that one was under some (fairly light!) cognitive load, the other was not. The Spinozan would hypothesize that the group under load wouldn't have the cognitive resources available to negate the propositions they tokened. The Cartesian view makes no such prediction and thus misses the effect that cognitive load has on belief formation.

An additional datum worth mentioning: in a second run of the experiment participants were also asked, after the video, to recall whether some particular statement was true or false

of the defendants. If the participants were burdened then they were significantly more likely to recall the false statements as true than the unburdened participants (they did so about 20% of the time), but were no more likely to recall true claims as false (in fact they were slightly more inclined to recall true claims as true than were the unburdened participants). The Cartesian view predicts that the burden should affect judgments of truth and falsity equally (because the system is being impeded before a judgment can be made) but that is not what we find. The bias is only to call false statements true, as the Spinozan view would predict.

Before we move on to the next killer raindrop, it is worth stressing that the effects in the study are not just effects on memory (the participants aren't just parroting responses), but are parasitic on the participants *believing* the lies they read and having the effects of their beliefs ripple through their cognitive system. In the first part of the study the participants not only processed the lies fed to them, but they made (presumably unconscious) inferences from those beliefs which then informed their judgments of the duration of the sentence and the character's likeability. This is quite interesting because it shows that the false information that is acquired acts like a belief in a hitherto unseen way: the information is informationally promiscuous, a hallmark of beliefs. Informational promiscuity has been previously suggested as a criterion for separating beliefs from other belief-like, sub-doxastic states (like intramodular representational states, e.g., the representations inside one's language module; see Stich 1978). The beliefs the participants formed infiltrated and interacted with (presumably some subset) of their web of belief in order to produce the behavior the experiment detected.[31]

2.5.2.2. Passive Belief and Personality Assessment

[31] Note that if these states were just 'aliefs' such inferential promiscuity would not occur. For more on aliefs please see section 4.5.

In a slight variation on a theme evidenced in 2.5.2.1, participants showed similar effects of automatic belief acquisition under duress when assessing the likeability of fictional characters. Participants read brief biographies of two fictional people, Bob and Jack. The biographies consisted of twenty-nine sentences about both characters, four positive ones (e.g., "Bob fed the stray cat by his house"), four negative ones ("e.g., Bob fidgeted a lot during class"), and nineteen neutral ones (e.g., "Bob enjoyed Mexican food"). Participants were told to learn these facts as well as they could. During the next phase of the experiment, participants received statements about Bob (though not Jack, for reasons that will become clear below). Participants were told to either assess the veracity of (the non-load inducing condition), or to speed read (the load inducing condition) the statements. They were also told that any novel statements they read during this phase were false.[32] Participants were broken up into two groups: one group would assess more valenced statements than they speed read, the other group would speed read more valenced statements than they assessed. These groups were then further bifurcated based on whether they were speed reading (or, conversely, assessing) mostly positive or mostly negative sentences. For example, one group assessed twenty likeable sentences (sentences that made the character seem likeable) and speed read only two dislikeable sentences, while another group speed read twenty dislikeable sentences and assessed only two likeable sentences. The group that speed read a majority of positive claims rated Bob as much more likeable than did the group that assessed a majority of positive claims, even though both groups were presented with an equal number of true positive sentences. The same pattern held for negative sentences: the groups that speed read a majority of negative sentences formed a more negative impression than the group that

[32] None of the novel statements contradicted the information they previously received. This was done to control for orthogonal associative effects.

assessed a majority of negative sentences. Speed reading the positive (or negative) sentences affected one's opinion in a way that assessing sentences did not.[33]

It seems that the group that was assessing was able to reject the false valenced claims, but the group that was speed reading was not able to reject the false valenced claims because of the cognitive load. This is because the assessing groups had the requisite time and cognitive energy to allow for the rejection process to run its course. Moreover, mere principles of association alone couldn't explain these data: before being asked the final likeability questions about Bob, participants were asked the likeability questions about Jack (who, was introduced in the biography section, but who the participants never read any false claims about). There was no difference in Jack's likeability between any of the groups. Thus, it couldn't be that participants were just primed to think (e.g.,) positive things about the character; if they were, then Jack's ratings would have been significantly affected. Rather, it seems that the participants believed the false statements that they speed read.

The morals of this experiment are the same as the one above: 1) there is an asymmetry in processing true vs. false claims that is missed by the Cartesian theory, though accounted for by the Spinozan one, and 2) this asymmetry is based on belief acquisition. The sentences that were speed read had to be integrated with other information in order to produce the final likeability judgment and such integration is a paradigmatic feature of belief.[34]

2.5.3 The Impotence of Knowing What Is False Before Encountering It

[33] Gilbert et al. 1993.

[34] Which is not to say that beliefs aren't often 'fragmented '(Egan 2008). I suppose that most beliefs are inferentially promiscuous to a degree, but don't actually interact with one's whole web of belief. Frankly, I bet most people's beliefs are highly fragmented and kept in context specific stores to facilitate not just further beliefs, but also to buttress one's psychological well-being (more on this in section 5.2.2).

One would think that if you knew that you were about to encounter false information and, as the Cartesian theory supposes, you had the ability to withhold assent, then you would not form beliefs based on the false information you subsequently encounter. However, the next two studies deal with situations in which people know that they are about to encounter falsehoods of certain sorts and yet still can't help but form beliefs based on the falsehoods. Which, to beat a dead horse, is just what the Spinozan, but not the Cartesian, would predict.

2.5.3.1 Belief Perseverance in the Face of Debriefing and Prebriefing

Another telling set of experiments comes from the literature on belief perseverance in the face of experimental debriefing. In a typical experiment, an experimenter asks participants to read a bunch of suicide notes and to sort the real ones from the fakes. In Ross et al. (1975), participants encountered twenty-five pairs of notes and were told that one note from each pair was a real note, one note a fake. After seeing each pair participants would judge which note was real and which fake and were then given feedback on their performance. After receiving the feedback the participants were (partially) debriefed. During the debriefing the participants were told that all the feedback they received was fictitious, it being arbitrarily determined beforehand regardless of the participants' responses. After the debriefing the participants were asked to estimate both how many times they actually answered correctly and how many correct answers an average person would give. Sadly, the information in the debriefing session did not affect participants' opinions about their ability: if the participant originally received positive false feedback (e.g., twenty-four out of twenty-five correct) they believed that they were better than average at the task, and if they received negative false feedback (e.g., seven out of twenty-five correct) they believed they were worse than average at picking out real suicide notes from fake ones.

The aforementioned experiment is generally not taken to illuminate anything about belief acquisition per se. It seems that the participants formed their beliefs in a reasonable enough way, based on the experimental feedback. Once they are told that the feedback was non-veridical they may just have trouble updating their beliefs. Perhaps beliefs are 'sticky,' in that once one has a belief, that belief is hard to get rid of. If so, then the debriefing effect wouldn't tell us about anything belief acquisition per se, but rather belief perseverance.

But what happens if the people are briefed before they take part in the study and receive false feedback (call such a technique 'prebriefing')? What if before sorting the notes they are told that the feedback they are about to receive is bogus? The Cartesian view predicts that if we tell people beforehand that what they are about to read is false, and they have no reason to distrust what we tell them, then, ceteris paribus, they will approach the stimuli skeptically, withholding forming any beliefs about their ability if those beliefs are based on the bogus data. On the other hand, the Spinozan view predicts that since people believe everything they token, they'll be stuck believing propositions that they encounter even if they know beforehand that they are false.

As predicted by the Spinozan view, but not the Cartesian view, prebriefing the participants beforehand does not impact the participants' judgments about their ability. Wegner et al. (1985) replicated the Ross study except the participants were told *prior* to the task that the notes and the feedback are dubious. Yet even after the explicit prebriefing the participants continued to behave as if the feedback was veridical. They were unable to reject the feedback they received, even though they knew the feedback was bogus. These perseverance effects are easily explicable on the Spinozan view: the knowledge of the feedback persists because the participants automatically believe the feedback when they hear

it, even though they know the feedback is false. The prebriefing effect helps to verify the first and third properties of the Spinozan theory. Equally importantly, these persistence effects are anomalous on the Cartesian theory, particularly casting doubt on the first property of the view. If the Cartesian theory were correct, then people would have the ability to withhold assent from propositions they encounter. Surely, if there was any time that this ability would rear its head, it would be in a situation where you know that you are about to encounter false information. Thus, the study just detailed gives us a perfect chance to allow the participants ability to shine, yet the people just could not reject the information they encountered. Such data should make one very wary of the Cartesian theory.

2.5.3.2 Never Trust a Fake Smile…At Least When You Know It's Fake Before Seeing It

Participants were presented with a series of smiling faces. They were told either before or after the presentation of the face that the face was either a fake smile (because the person had previously had a bad mood induced before being asked to smile) or a real smile (because the person has previously had a good mood induced before being asked to smile). Participants were also told that they would sometimes hear a tone and they'd have to push a button indicating whether the tone they heard was high or low pitched. During this phase, subjects were asked to remember which faces contained real smiles and which ones contained fake smiles. After the learning phase, subjects were re-presented with eight faces they had seen, four of which had been previously shown with the interrupting tone. In the test phase, subjects identified real smiles correctly 60% of the time when the smiles were originally presented uninterrupted and 65% of the time when originally presented interrupted (an insignificant difference). However, the correct identification of fake smiles was at 65% when originally presented uninterrupted and only 35% when interrupted (a significant

difference). The results were the same whether or not the participants were told that the smiles were fake *before* or after they saw the picture.[35] Thus, even when subjects knew that they were about to see a false smile they couldn't help but see it as a real smile.

The moral to draw here is similar to the moral to be drawn from the studies mentioned above. If the Cartesian theory were right, then we'd expect people to show no asymmetry in memory between remembering truths and remembering falsehoods. But once again we see that cognitive load sheds light on a deep-seated asymmetry between truths and falsehoods: being cognitively busy during the learning situation causes people to misremember falsehoods as truths, but not truths as falsehoods. The study mentioned here is just more evidence toward showing that the time at which one learns that they are acquiring a falsehood matters not to the asymmetry in the way we process truths and falsehoods. This asymmetry is missed by the Cartesian theory, but not the Spinozan one.

2.5.4 Personality Metrics

Evidence for the Spinozan theory comes from a wide array of sources. Another example comes from an unlikely place: personality psychology. When studying personality psychology, researchers often present subjects with a list of personality attributes and ask participants to evaluate how much the attribute describes their personality. Consider a personality survey where participants are given twenty statements and are asked to answer, for each statement, whether the statement applies to them or not. The participant answers 'yes' when the statement applies to them and 'no' when it doesn't. On ten of the questions an answer of 'yes' corresponds with being an introvert and on the other ten questions an answer of "yes" corresponds with being an extrovert. On such a scale a 'perfect' introvert would be

[35] Gilbert et al. (1990).

one who answered "yes" to the ten introversion questions and "no" to the ten extroversion questions, while a perfect extrovert would reverse the answers. When using such methods researchers have found that their data are sometimes compromised by 'yea-sayers,' i.e., people who are apt to respond affirmatively to whatever question they are asked. For example, a perfect yea-sayer would respond to the aforementioned study by answering "yes" to all twenty questions, thus confounding the personality metric.[36] The perfect 'nay-sayer' would reverse the pattern of the perfect yea-sayer.

If negations are processed subsequent to affirmations, as the Spinozan view would have it, then we should expect that nay-saying takes more energy, and thus time, than yea-saying. This is because, for the Spinozan, the first stage of encoding/accepting is passive and effortless whereas the second stage of rejecting is active and effortful. Thus, the Spinozan nay-sayer would have to first encode the property as applying to them and would then have to go back and reject the property, whereas the acquiescing yea-sayer would just need to passively encode the property. Additionally, the Spinozan view predicts that if people are put under cognitive load while answering one of these personality metrics, then yea-saying should increase relative to an administered personality metric that lacks any load-inducing element. This is because the load makes the participant more cognitively enervated and therefore less able to summon the energy to reject the proposition. In contrast, the Cartesian symmetrist position predicts that because accepting and rejecting are products of the same underlying process, yea-sayers should take the same amount of time as nay-sayers and both should be equally affected by cognitive load.

[36] These acquiescent yea-sayers often confound unbalanced personality scales, like the Minnesota Multiphasic Personality Inventory (see Block 1965) and the California F-scale (see Couch and Keniston 1960).

Both Spinozan predictions were borne out in Knowles and Condon (1999). Participants received a counterbalanced 100 item personality questionnaire and had their reaction times measured. Yea-sayers were operationalized as those who answered affirmatively on fifty-three or more of the items, and nay-sayers as those who answered affirmatively on forty-seven or fewer of the items. The middle group counted as appropriate responders. The response times for yea-sayers were significantly quicker than the response times for either of the other two groups. In fact, when we look closer we can see that the response patterns perfectly conforms to the Spinozan hypothesis, with yea-sayers taking longer than appropriate responders, who in turn took longer than nay-sayers. This response pattern is directly at odds with the third Cartesian prediction.

Cognitive load also affects yea-saying in the way predicted by the Spinozan, but not Cartesian, hypothesis. In a related study participants were split into two groups, both of which were asked to answer twenty counterbalanced personality questions. Intermittent music was playing in the background for both sets of participants. One set of participants was put under cognitive load by listening to musical notes and attempting to distinguish notes that came from the piano from those that came from other instruments. The non-loaded group heard the same sounds but wasn't asked to attend to them. The group under load was significantly more apt to answer affirmatively to the questionnaire, thus confirming the second Spinozan prediction and disconfirming the second Cartesian prediction.[37] A theory that sees acceptance as passive and automatic but rejection as active and effortful, as the

[37] This replicated a similar findings reported both in Mcgee 1967 and Trott and Jackson 1967. In those studies the load was of the fast-response variety (e.g. respond to every question in three seconds or less) and thus was not continual throughout the task as in the study reported. The findings were uncovered even without continual load.

Spinozan theory does, predicts that load affects nay-saying differently than yea-saying because only the former is active and effortful, thus only the former is a viable candidate to be affected by load. A theory that sees acceptance and rejection as part of the same underlying active mental process, as the Cartesian theory does, cannot explain such findings.

2.5.5 Counter-attitudinal Communications

In Festinger and Maccoby (1964), experimenters showed participants films that contained anti-fraternity messages. The films had a speaker who argued that fraternities should be abolished because they negatively impacted universities by breeding dishonesty, racial prejudice, and social snobbery into college life. The participants were college students (from sophomore on up) who were either fraternity brothers or independents (i.e. non-fraternity related). Both fraternity brothers and independents were further split into two conditions, a cognitively burdened one and an unburdened one. The unburdened group watched a film that had a man plainly delivering the anti-fraternity brother speech. The burdened group heard the same speech, but instead of just showing a man plainly delivering the speech, the background to their film was a highly incongruous, distracting, and amusing film that had sound effects and music, but no talking. After watching the film the participants were asked to rate how highly they thought of fraternities along a number of dimensions.

The fraternity brothers, who were unsurprisingly pro-fraternity in their antecedent attitudes, were not expected to be too happy with the film's message. Likewise, the independents were disposed to come into the experiment harboring more negative opinions to fraternity lifestyle. However, the overall opinions of these subjects aren't what one should keep their eye on; rather, what interests us is how the distracting (i.e., cognitively burdensome) film affected the participants. The fraternity brothers in the burdened condition

were more apt to be persuaded by the anti-fraternity message than their non-burdened brethren.[38] The Spinozan theory predicts this asymmetry, for it predicts that since rejecting a message is an effortful endeavor, one can only reject a message when they are unburdened (and thus have the requisite cognitive energy).[39] In contrast, it would be quite natural for the Cartesian theorist to expect distraction to have the opposite effect. If assenting (and dissenting) are both active processes (as per the second Cartesian prediction), then one would expect that being distracted would shut down either active endeavor and make people less likely to change their opinion in any direction, never mind in a counter-attitudinal direction.

More evidence in favor of the Spinozan hypothesis comes from the independents. Since the independent group antecedently agreed with the speaker's message, the message did not differentially affect the two independent groups. This datum is explicable as follows: since the independents didn't need to argue against the message, they didn't need the extra cognitive effort that would be drained in the unburdened condition. Thus, the two conditions shouldn't differentially affect the independents for all they were disposed to do was passively agree with the message anyway.

The upshot of this discussion is that the Cartesian view cannot explain why participants who hear counter-attitudinal communications while under cognitive load are

[38] This isn't even a strong enough statement of what transpired. The fraternity brothers in the burdened condition weren't just *more* apt to be persuaded than their non-burdened counterparts, rather the burdened brothers were apt to be persuaded by the message tout court.

[39] One may be apt to argue that because the distracting film contained an entertaining background the distracted fraternity brothers were positively reinforced and hence they were more apt to think positively of the message. However, the participants' comments make this interpretation highly doubtful. Fraternity brothers in the burdened condition complained about having to hear the message and thus not being able to focus on the entertaining film. They often claimed that the incongruity between the two was distracting (e.g. one participant wrote, "I could not see any tie in between what was being said and what was being shown. It was very hard to concentrate on what was being said without completely looking away from the movie" [Festinger and Maccoby ibid, p. 366]).

much more likely to believe the counter-attitudinal propositions, while the Spinozan theory has the requisite resources to explain such cases. Here, as elsewhere, the effects of distraction on belief acquisition and attitude adjustment are inexplicable on the Cartesian hypothesis, though predictable on the Spinozan.

2.5.6 Perception, Attribution, and Automatic Belief Uptake

Our tour continues with an overview of some studies showing automatic belief acquisition in traditional 'fundamental attribution error' (Ross, 1977, or equivalently, 'correspondence bias,' Jones, 1987) cases. Both of the cases below involve reflexive beliefs being formed from perceptual processes and eventuating in misattributions.

2.5.6.1 Explaining Others Behaviors

In a study testing people's folk attributions of dispositional versus situational inferences, Gilbert (2002) showed participants silent videos of a woman being interviewed. Although the subjects could not hear the interview, they were told what topics were discussed. The videos were classified into two groups, a "sadness condition" and a "happiness condition." The sad videos contained interviews where the actress was asked (wait for it…) sadness-inducing questions about her life (e.g. "Describe a time when your parents made you feel unloved") and in the happiness-inducing condition the actress was asked happiness-inducing questions (e.g. "What is the nicest thing your parents have ever done for you?"). The participants viewed the videos in a booth that had a camera on top of the monitor, pointed at the subjects. They were told that the camera was an eye tracking device (a "parafoveal optiscope", which sounds very fancy, but is just made-up). Half the subjects from each group were put into an unburdened condition and half into a burdened condition. The unburdened condition subjects were told that a series of words would appear

and disappear on the screen and that these words could be ignored because they were tangential to their experiment. In the burdened condition the subjects were just told that they could not look at the words, for if they broke eye contact from the actress and looked at the words the camera would stop working and the experiment would not produce any reliable data. Thus, the unburdened subjects were told they *could* (but needn't) ignore the words, whereas the burdened subjects were told they *must* ignore the words.[40]

This mere act of self-regulation was cognitively demanding enough to make a noteworthy difference in the participant's responses. At the end of the study, when subjects were asked how (dispositionally) happy or unhappy the actress was, those in the unburdened condition were apt to account for the situational constraints whereas those in the burdened condition did not. The burdened subjects believed that the actress was disposed to always be happy (or sad, depending on what video they saw), whereas the unburdened subjects realized that they knew little about the subjects dispositional state. The most natural way of describing this case is that participants in both conditions reflexively believed what they saw and that the participants who were not faced with an increased cognitive load had the ability to correct for (i.e., reject) their initial impressions, thereby rejecting their initial beliefs. Participants faced with an increased cognitive load, however, could not correct for their initial perceptions because they were cognitively burdened. Once again, the impact of load

[40] Note that the mere self-regulation of behavior (e.g. being told not to look at something) is enough of a burden to induce cognitive load. Think about any social situation—can you imagine one where the majority of people aren't self-regulating (particularly impeding) some form of response? Note also how light this cognitive load is throughout these subjects. These subjects aren't being asked to also do calculus or count backwards by fours from 1,000, they are just asked to push a button when a tone arises, or when the number 5 shows up on a slow moving crawl. This type of load is exactly the type of load that we should expect that people are constantly being put under in real life, ecologically valid situations. When I'm walking down the street I'm monitoring the street for taxi-cabs that might come careening at me, so I've the requisite cognitive load at play, as I pass by and read the billboard advertisement. It's thus no wonder that when you're a guest sitting at the high table, the Don sounds exceptionally persuasive.

on people's perceptions serves to uncover an asymmetry in how people accept and reject information and this asymmetry is anathema to the Cartesian theory, while at home in the Spinozan.

2.5.6.2 Explaining Your Own Behavior

As opposed to the study discussed directly above, which was a study of assessing other's psychological states, a similar anti-Cartesian effect can be seen in studies of assessing one's own mental states. In Gill et al. (1999) participants listened to music that was designed to either depress or elevate one's mood. The participants were then given forty-four adjectives and asked to rate which ones accurately described their personalities (and not their transient mood). Keeping to the script, participants were split up into either a hurried condition or an unhurried condition. Those in the hurried condition were asked to respond as quickly as they could (thus inducing the appropriate cognitive load), whereas those in the unhurried condition were asked to take their time and reflect on their answers. Unsurprisingly, those in the hurried condition drew dispositional inferences based on their current moods, whereas those in the unhurried condition corrected for the situational constraints (and didn't let the music they were listening to prime their answers).[41]

Note that the dispositional inference is the inference that is based on merely taking what you perceive as true. The Spinozan would explain the effect by nothing that since the participants clearly perceive (and thus token a mental representation of) their own behavior, but don't token a corresponding mental representation about the situational inference, they

[41] This is an interesting datum in favor of the Rylean, or more generally behaviorist, view of self-knowledge (Ryle 1949). Ryle suggested that people draw inferences about themselves in the same way they draw inferences about others: through observing their own behavior. The data reported suggests that people use the same mental processes in order to make folk-psychological judgments regardless of whether they are judging their own folk-psychological states or others, for cognitive duress affects judgments and perceptions of others exactly as it affects judgments and perceptions of one's self.

end up believing what they token and not adjusting their beliefs in light of the relevant evidence. Contrarily, the Cartesian cannot account for why the fundamental attribution error (which is of course an error of attribution based on mistaken belief) is exacerbated by cognitive load.

2.6 Conclusion

Before concluding this chapter, I'm going to cite some plain, but sagacious advice about how to compare competing hypotheses. When discussing competing theories of propositional attitudes Fodor once wrote "It's the mark of a bad theory that it makes the data look fortuitous… we should prefer a theory that explains the facts to one that merely shrugs its shoulders" (Fodor 1981a, 180-181). The Cartesian theory misses a slew of evidence that it should be able to account for, but can't. The theory supposes that one can withhold assent, yet time and again no evidence is garnered in defense of this ability. Perhaps the Cartesian theorist can go datum by datum waving his hands and creating exceptions, but even if she could (which is doubtful, see chapter 4), from the viewpoint of the theory as stated all of the evidence marshaled here is simply inexplicable on her theory. At best the Cartesian theory has no explanation to offer; at worst, it's consistently facing evidence that refutes the theory.

The conclusions from our painstaking tour should be clear enough: there are some strong reasons to be skeptical of the Cartesian theory of belief acquisition. The evidence that we've encountered so far doesn't just tell against the Cartesian theory, it also provides support for the Spinozan model. However, I think the Spinozan theory has more going for it than just explaining the observations we've canvassed so far. In the next chapter, I will give an abductive argument for why one should believe that the Spinozan theory is a reasonable research program, if not an actually accurate representation of the structure of belief fixation.

Then, in chapter 4, I'll survey a bevy of objections to the view and then end by taking stock of what repercussions the Spinozan theory has for reconceptualizing our picture of the architecture of the mind outside of belief fixation. But first: pudding.

Chapter 3: The Explanatory Capabilities of the Spinozan Theory: The Pudding

The previous chapter perused a bevy of experiments that should make one quite pessimistic of the prospects of the Cartesian theory while hopefully increasing one's confidence in the viability of the Spinozan theory. However, the skeptical reader may reasonably want to see that the Spinozan theory can do more than just conform to the extant data on how people deal with impinging information under cognitive duress. In this chapter, I will display the breath of the Spinozan theory. The theory's explanatory power, theoretical elegance, and generality give one yet another reason to trust the theory. I will argue that the Spinozan theory offers us a certain type of consilience with other psychological phenomenon outside of the somewhat parochial realm of the evidence we've encountered so far.

The Spinozan theory can help give a unified explanation of what prima facie appears to be a bunch of disconnected and problematic, if not downright mysterious, phenomena across philosophy and cognitive science. I take it that our options are as follows: either the Spinozan theory is true and we can explain a plethora of hitherto poorly understood phenomena or we reject the Spinozan theory because it doesn't comport with our intuitions, in which case poorly understood phenomena remain poorly understood and we still have no grasp on the mechanisms of belief fixation. Since I take it that it's common ground that some explanation is better than none, I think the arguments contained in this chapter give us strong reason to take the Spinozan theory seriously, even if it strikes our ears as unintuitive. Of

course, the proof is in the pudding; the argument will only work in so far as the explanations given below are actually illuminating. Well then, to the pudding.

3.1 The Fundamental Attribution Error

The canonical formulation of the fundamental attribution error is formulated in terms of one's perceptions of others and the causal antecedents of their behavior. When we perceive an agent's actions we often fail to account for the situational constraints that cause the actions and instead explain the agent's actions by appealing to their personality traits. Thus, if you encounter me walking out of the hospital after staying up all-night with a sick child (or if you encounter me in a psychology experiment where I'm focused on being watched by an experimenter or if you stumble upon me sitting in class while I'm hoping that I look and sound professional,[42] etc.) you will be more apt to perceive me as a dispositionally anxious person, ignoring the particular situation you find me in. This example is not special: in general, behaviors which have situational factors as their crucial causal variable are instead misinterpreted as behaviors which display someone's basic character (hence the appropriate re-coining of the correspondence bias as the *fundamental* attribution error).

This canonical formulation of the fundamental attribution error is expected on the Spinozan view: since we believe what we token and we (ceteris paribus) token thoughts corresponding to what we perceive, we believe what we perceive. If you see that I am racked with anxiety, then you believe that I'm an anxious person. For you to override this belief, you'd need to reject it by allowing for the integration of additional information (viz. your knowledge of the operative situational constraints) to arise in your reasoning. However, as

[42] This has never happened, for I gave up all pretensions of appearing professional long ago.

we saw in section 2.5.6, when you are under cognitive load, you are less apt to be able to engage in such overriding for load shuts down our ability to reject information and hence

exacerbates the fundamental attribution error.[43] Moreover, as we saw in 2.5.6.1., cognitive load can be brought on by the mere self-regulation of one's own behavior (as a reminder, in the experiment discussed there the onset of load occurred by asking people to avert their gaze and not read a crawl on the bottom of a computer monitor). When we're in social situations, we're apt to self-regulate and thus we're often already under cognitive load just from the situation at hand (e.g., when you are speaking you may want to ensure that you don't fiddle with your hands, or pace back and forth, or look at one's chest, or look at one's deformed ear or the chocolate stain on one's sleeve, etc.). Thus, we're apt to believe our perceptions at face value because we're unable to account for the additional (and, more often than not, imperceptible) evidence.

The Spinozan theory can also help explain the less canonical formulation of the fundamental attribution error, which deals with one's own perception of the causes of one's own behavior (Jones 1977). Say you and I get into a fistfight. The first formulation of the error predicts that I'll be more apt to think that you hit me because (e.g.) you're a pugnacious rogue (as opposed to someone who has just been hit in the face and is acting out of self-defense). However, there is another formulation of the error that applies to our own behavior. The second formulation states that I'm apt to explain my behavior (when it's more negative than positive) in terms of situational constraints (e.g. I'm punching you because you are an aggressive malcontent, or because you deserve it for the time you wronged those wombats, etc.). This formulation is also expected on the Spinozan theory. When I am acting, what's

[43] It is reasonable to suppose that the real effect of load is on the rejection of information and not necessarily the integration of information. After all, it seems (at least introspectively) plausible that the integration of information occurs unconsciously and automatically. If this is right, then it might be reasonable to suppose that integrating new information can occur while under load. Of course, this inference is far from apodictic for I don't think we yet have reason to suppose that the integration of information is mandatory. Perhaps a more conservative bet would be that load impedes both our ability to reject information and our ability to integrate information, therefore doubly exacerbating the fundamental attribution error.

most salient to me are (some of) the situational constraints on my behavior. After all, perception looks outward—when I act, I see what I'm *re*acting to in my environment. I see that you look angry and conclude that I punched you because of the terrorizing look you had on your face. I then believe that this is the cause of my behavior because I perceived it to be so and I'm under considerable load (if social situations alone bring on load then affectively charged social situations bring on that much more load) so I can't readjust my perceptions to take into account my own personality quirks (e.g. I'm a pugnacious rogue who likes to hit people).

One may object to the above explanation, instead preferring a seemingly simpler explanation with a classier pedigree. Heider's (1958) suggestion that "behavior engulfs the field" was offered as an interpretation of the position of the observer who watches social interaction. Heider's idea was that behavior is so salient that its vividness swamps the observer's focus on situational factors so that the observer focuses on the actor at the detriment to the environment. This type of explanation of the fundamental attribution error has no need to refer to entities like belief at all—it can presumably do all of the explanatory work just by using the concept of attention. However, the Heiderian explanation cannot explain the wide scope of the fundamental attribution error. This is because the error arises in contexts where it doesn't make sense to apply the notion of 'behavior engulfing the field.' For example, as Quattrone (1982) has pointed out, the error arises in cases of forced essay writing. Imagine I have been asked to write an essay in favor of lifting embargoes off of Cuba. Suppose further that someone then presents a third party with my essay and tells them of the task demands that surrounded the creation of the essay. In this case people still infer

essay-consistent attitudes on behalf of the essay writer (even though the participants are told about the task demands).

Here the idea of the behavior engulfing the field does not seem applicable: focusing on the essay implies neither a focus on the agent nor the situation for both the agent and the situation are invisible and only the products of behavior remain. Furthermore, even if one liked the Heiderian idea and tried to pursue it as an explanation in the essay writing paradigms, it still wouldn't explain why people infer that the essay writer held essay-congruent attitudes as opposed to essay-incongruent attitudes. All the Heiderian principle tells us is that people will focus on behavior to the detriment of the rest of the scene; it does not help us figure out how to get from that fact to the fact that people infer essay-congruent attitudes (or likewise, that people have characters that are congruent with their behavior. Note that there is no reason in Heider's explanation that would rule out people drawing the inference that the agent is akratic or otherwise acting contrary to their character).

On the other hand, the Spinozan theory has no such trouble in explaining the generality and direction of the fundamental attribution error. The Spinozan Theory is just as applicable to the essay writing cases as it is to perceptual cases of the error. The generality of the Spinozan theory thus allows it to succeed where others have failed. Moreover, the Spinozan theory can also serve as a reductive theory: it allows us a way to understand this higher level social psychological phenomenon (the fundamental attribution error) in terms of a lower level architectural constraint (the Spinozan idea that thinking entails believing).

3.2 The 'Mere Possibilities' Version of the Confirmation Bias

The 'confirmation bias' refers to people's tendency to search for confirmatory, but not disconfirmatory, evidence for the hypotheses they antecedently believe (Klayman and Ha

1987). The bias is explicable on a basic dissonance theory (e.g. Festinger 1957). A potted explanation goes something like: if we harbor belief X and we find evidence that speaks against X, ascertaining such evidence will put us into a dissonant state. Since dissonant states feel bad (Zanna and Cooper 1974)[44] they act as negative reinforcers and through classical conditioning they reinforce us to not search for such disconfirming evidence. This type of explanation explains why the confirmation bias arises in cases of a previously held belief.

However, the ubiquitous 'mere possibilities' version of the confirmation bias arises in cases where people are merely considering a proposition and have not yet endorsed the proposition (Snyder and Swann, 1978; Snyder and Campbell 1980; Swann et al. 1982). For example, if I ask give you a set of objects and a rule that the objects are supposed to conform to and then ask you if the rule holds, in general you will search for objects that comport with the rule as opposed to objects that disconfirm the rule (Wason 1961, Wason and Johnson-Laird, 1972). A more specific example: if people are asked to consider if they are happy with their social life, they generally respond that they are, but when people are instead asked if they are unhappy with their social life they also generally respond that they are (Kunda et al. 1993). In these cases people search their memory for information that would confirm the question and then stop their search once they reach such information.

Dissonance theories falter when attempting to explain the mere possibilities formulation because (by Cartesian assumption) people aren't yet invested in thoughts that they merely entertain. Since they're not yet invested in these propositions their self-image is not hostage to veracity of the propositions, thus they shouldn't be apt to show classic

[44] Cognitive dissonance is caused by two cognitions that are inconsistent with each other, but being in a dissonant state is itself not a cognitive state per se, but a negative motivational state (Cooper 2007). N.b., it's notoriously difficult to spell out what exactly 'inconsistent' amounts to, but it's clearly not just strictly logical inconsistence.

dissonance avoidance strategies toward disconfirming evidence. As a consequence, the mere possibilities formulation of the confirmation bias is a standing mystery. Happily, the Spinozan theory can explain it without contorting itself merely by pointing out that propositions that one merely contemplates one automatically believes. Since one believes merely contemplated hypotheses, one is already invested in the hypotheses merely by considering them. This analysis allows the dissonance explanation to get a foothold and start doing its explanatory work, thus explaining this hitherto recalcitrant phenomenon.

3.3 Anchoring and Adjustment

'Anchoring and adjustment' has been used to mean different things. Most confusingly, it has been used to refer both to the experimental *anchoring procedure* whereby one gives a salient and uninformative numerical value to participants before they have to make some numerical judgment and it has also been used to refer to a *mental process* that is active in the experimental procedure. It is the latter that will be of interest to the current discussion. The Spinozan theory can help explain the mental process that is at work in the anchoring and adjustment experimental paradigm. Doing so would be an explanatory coup, for not only is the effect robust and the process behind it mysterious, but also because theorists have taken the anchoring and adjustment effect to have widespread repercussions outside of the numerical anchoring and adjustment paradigm. Whatever it is that is supposed to explain the numerical form of anchoring and adjustment has also been presumed to explain a range of effects like egocentric perspective taking in language production (Keysar and Barr

2002),[45] base-rate neglect (Tversky and Griffin 1992),[46] overconfidence (Tversky and Griffin ibid),[47] and the above (and below) average effect (Kruger 1999).[48] It is not my goal to argue that these other effects are actually caused by the same mental process underwriting the numerical anchoring and adjustment paradigm. I find extending the anchoring and adjustment explanation to these other phenomena to be a tantalizing possibility, but I'm

[45] The idea here is that one anchors on one's own mental states and then adjusts to other's mental states. So, for example, when uttering an ambiguous phrase the speaker first thinks that the phrase is unambiguous (because she is anchored on her own intentions) and then (generally insufficiently) adjusts to try and take into account the hearer's knowledge of the speaker's background intentions. The effect arises even as a hearer: if a hearer has additional information that she knows another observer doesn't have, she will still anchor on the information available to her and then (generally insufficiently) try to take into account other observer's knowledge and adjust accordingly (Keysar 1993, 1994).

[46] Tversky and Griffin hypothesize that people anchor on the salient problem at hand and then adjust to the base rates (as opposed to just taking the base rates into consideration up front). For example, imagine I told participants that I interviewed 100 people, ninety of which were artists and ten of which were engineers and then I give them a small transcript of five of the interviews. If in these transcripts the people come off as boring, mathematically inclined folks, participants will anchor on the vividness of transcripts and ignore the base rates even though they know them. This would lead the participants to, in general, guess that the interviewees were engineers and not artists.

[47] Participants anchor on the strength of the evidence (e.g., how glowing a letter writers recommendation is) and then adjust to the weight (e.g., how credible the letter writer is). Thus, super strong, glowing letters of recommendation often swamp a letter reader's impression even when the letter reader knows that the letter writer isn't a credible writer (Tversky and Griffin ibid.). Consequently, it may make sense to include a buddy's letter of recommendation in one's dossier even if the selection committee knows full well that the buddy is not a credible letter writer.
(Side anecdote on a similar point: during my interview for the Princeton Society of Fellows, I had to also interview with Princeton's philosophy department, during which I had the pleasure of spending a few hours with Gil Harman. Gil was perplexed as to why he was talking to me. He explained that because the interview situation was fertile breeding ground for cognitive biases, the department had decided to eschew interviews with job candidates [the thought being that a bad interview could easily swamp a candidate that looks great on paper]. Gil reasoned that the interview was subject to all sorts of influences which were not really important in assessing a candidate, and that instead the candidate's CV was much more indicative of the candidate's value [the CV acting as the base rate here and the interview acting as the vivid example which often trounces the base rate for 'no good reason']. I then asked Gil if the selection committee reads letters of recommendation and takes them seriously. Unsurprisingly, they do. I then proceeded to ask why they'd use one flawed indicator and not the other, since both interviews and letters are subject to the same anchoring and adjustment biases. He responded by saying that they have to use some indicators other than the CV. I then retorted that aren't more flawed indicators better than fewer? After all, we use the GRE and GPA as indicators to get into graduate school even while knowing that they're flawed indicators. I then proceeded to not get an offer from Princeton [not that this was Gil's doing—he was perfectly lovely and willing to discount the interview anyway.])

[48] The 'above average effect' is the name for people's tendency to think they are above average in tasks where absolute skills tend to be high (e.g. driving), and below average in tasks where absolute skills tend to be low (e.g. juggling). For more on this effect and its possible relation to anchoring and adjustment see footnote 57.

ambivalent as to whether this program can be carried out. As a consequence, the main text will just concern itself with explaining the basic numerical anchoring and adjustment effect; at the end of the anchoring and adjustment section I will explain my hesitation toward these extensions in a footnote (footnote 57). If these other phenomena do turn out to be manifestations of the same underlying process as the anchoring and adjustment effect, then the Spinozan theory's explanatory power is that much stronger. So, although I am neutral as to whether these other effects are caused by the same process that creates the anchoring and adjustment effects, I welcome these suggestions.

3.3.1. The Numerical Anchoring and Adjustment Effect

In the prototypical anchoring and adjustment paradigm (e.g., Kahneman and Tversky 1974) experimenters ask participants to figure out numerical values for some arbitrary questions, like 'How old was Gandhi when he died?',[49] 'What is the freezing point of vodka?',[50] 'When was George Washington elected president?'[51], What percentage of the UN is made up of African nations?'.[52] Before participants are allowed to answer the target question, the experimenter arbitrarily selects a number (e.g. by spinning a wheel, or by using the participants social security number, or by a randomly chosen card, etc.) which serves as an 'anchor.' Importantly, the participants are made aware that the number selected is indeed arbitrarily selected (by, e.g., making the participants spin the wheel or choose the random card). Participants are then asked whether the answer to the target question is higher or lower

[49] Gandhi died at 78.

[50] For 80 proof vodka it's approximately -16.51 Fahrenheit.

[51] He was elected in 1779.

[52] -15. It's surprising but true. You can look it up.

than the arbitrarily picked number. (Say the number that came up on the wheel was 1776. Participants would then be asked, e.g., whether George Washington was elected president before or after 1776.) After answering this question, participants are then allowed to give an exact answer to the original question. The randomly generated anchors make a significant impact on the subjects' answers.[53] For example, people will guess that Gandhi died at 50 if they first had to decide whether he died before or after he was 9, and they'll think he died at 67 if they receive 140 as the anchor (Strack and Mussweiler 1997).

Explanations of the anchoring and adjustment effect are scant at best. For example, the traditional 'explanation' of the effect is that people anchor on a value and then adjust up or down from that value (Kahneman and Tversky 1974). This explanation is just a restatement of the phenomenon (not that the authors didn't realize this—as they mention, they just didn't have any other explanation on offer). Since then, the main explanation of the effect is that it is produced by "increased accessibility of anchor-consistent information" (Epley and Gilovich 2001, Mussweiler and Strack 1999, 2000). Although this may seem like a novel explanation, it is just an instance of a broader trend, the bias towards searching for confirmatory evidence, the confirmation bias. Hence, the confirmation bias is supposed to explain the anchoring and adjustment effect. But as we've just seen in section 3.2, the confirmation bias itself presupposes the Spinozan theory. The confirmation bias is only supposed to be in play when we are searching for evidence to confirm an *already* held belief, so by accepting the confirmation bias explanation the non-Spinozan theorist just doubles her mysteries, for she'd also need to explain why merely contemplated hypotheses are believed. But the Spinozan theory can alleviate these mysteries. Anchoring and adjustment effects

[53] The participants generally move halfway towards the anchors as compared to participants that don't encounter an anchor (Jacowitz and Kahneman 1995).

arise because participants believe that the anchor they're given is actually the answer to the question they're posed. Participants believe that the anchors are the correct answer because they merely consider that possibility, and consideration causes belief.[54]

One may object to this explanation on the grounds that anchor consistent information becomes more available not because of the confirmation bias, but instead because of mere semantic priming. Maybe one's 'accumulator' (see Gallistel and Gelman 1992) is active when the number 140 comes up, and this primes other closer numbers. Or perhaps it's one's symbol for 140 that arises and primes other closer numbers. However, there are two reasons to discard this objection. For one, if the numerical priming story were true, then we'd expect what participants do when they adjust is to continually slide along the number line until they reach a limit, one presumably dictated by the extent of the priming effect.[55] However, the one study I know of which has attempted to test whether the adjustment process is serial or continuous (like the priming story would suppose), Epley and Gilovich (2001), have data which speaks against the priming/sliding view and propose instead that the adjustment phase is a series of jumps, which are "discreet minadjustments coupled with hypothesis tests" (Epley et al. 2004, p. 328). Such jumps do not seem to be explicable on the priming hypothesis.

---

[54] Here, as elsewhere, lies a tacit ceteris paribus clause. When participants are asked a question and then given the anchor the participants must form a thought that turns the interrogative into a declarative. Presumably, most participants do this automatically. If participants did not make such a transformation then the given explanation wouldn't hold (and presumably the participants' answers wouldn't show the anchoring and adjustment effect. Evidence for this claim follows immediately below in the discussion of the Chapman and Johnson.).

[55] This has been shown to be the process by which numerical priming works (see, e.g., Dehaene 1997), thus explaining why size and distance effects arise in numerical priming experiments (Moyer and Landauer 1967; Meck and Church 1983).

More direct evidence against the priming explanation comes from Chapman and Johnson (2002). They point out that not any old numerical value will cause the anchoring and adjustment effect—rather the arbitrary numerical value has to be considered relevantly related to the question at hand, even if it's a grossly unreasonable value. For example, say I asked you to consider what the average income of a New Yorker is. Before you answer this question, I ask you to spin the wheel to generate an anchor and the wheel comes up with the number 90,000. Now here's the important part: if the intermediate question is orthogonal to the original question, the anchoring and adjustment phenomenon will not arise. Suppose after the wheel spits out 90,000 I ask you if 90,000 is greater or less than the square footage of Buckingham Palace and you answer in whatever way you deem correct.[56] Suppose further that after answering this comparison question we return to the original question of the average income of a New Yorker. In this case the arbitrarily derived anchor will not affect your subsequent judgment. The anchors only affect your judgments when the anchors are understood as related, even if unreasonably so, to the question that you are considering. In other words, you have to complete a thought that involves the anchor as the answer to the question you are considering. Of course, this is decidedly not the way that priming works. Priming works by mere associative activations (or 'construct activations') spreading through one's cognitive system. So, if the anchor was just working as a prime we'd expect it to work regardless of what question a person is considering, however that is not how the anchoring and adjustment paradigm works, so we can be satisfied that priming is not the explanation of the anchoring and adjustment effect. The only contender explanation is the Spinozan theory, thus lending strong evidence in favor of theory. The anchoring and adjustment effect is as

[56] Buckingham Palace is 828,818 square feet. In stark contrast, my current apartment is 350 square feet.

robust an effect as we find across the literature. A theory which is independently plausible and can explain this effect should be taken quite seriously.

It is worth noting that I'm not the first to propose that the anchoring and adjustment effect arises because participants believe that the anchor is the answer to the question they're considering. Jacowitz and Kahneman (1995) also floated this explanation. However, they did so not for architectural reasons, like the Spinozan theory does, but instead for Gricean pragmatic reasons. Jacowitz and Kahneman suppose that participants' reason that experimenters wouldn't give them the anchors if the anchors were *truly* irrelevant. Thus, Jacowitz and Kahneman suppose that participants believe that the anchors are good answers for (broadly speaking) *rational* reasons. This seems utterly implausible. If the subjects are going to end up believing that Gandhi lived to 140, our principles of charity should dictate that they don't reason their way there. When subjects see that the anchors are created from their (e.g.) social security number, do we really want to say that they then surmise that this is the correct answer? After all, it is often the participants themselves who (e.g.,) spin the wheel of fortune thus seeing exactly how arbitrarily derived the anchor actually is. By supposing that participants reason their way to believing that the anchor is (more or less) the correct answer we attribute to the participants not just a smidgen of irrationality (which is often reasonable to do) but instead massive stupidity. People may be unreasonable in all sorts of circumstances, but if we grant them Jacowitz and Kahneman's suggested patterns of reasoning, we'd have to suppose that they are absolute idiots, a conclusion which strikes even my cynical ears as implausible.

A more reasonable supposition is that subjects, in the first instance, don't have a choice about what they believe. They end up believing that Gandhi lived to 140 because they

are forced to believe so by the architectural set up of the mind and not because they follow some very suspicious logic based on a leap of faith in the experimenters' intentions (and skill at deriving a correct answer from a subject's license!). In sum, it is reasonable to conclude that the Spinozan theory is the only reasonable explanation of the vast anchoring and adjustment data.[57]

[57] As advertised earlier, I will now return to the discussion about extending the anchoring and adjustment explanations to cover seemingly unrelated phenomena. The following case study should suffice to show the source of some of my hesitation in endorsing these extensions. Let's focus on the above average bias. The anchoring and adjustment explanation of the above average bias states that when people are asked how good they are at a task compared to the population in general they first anchor on their own ability at performing the task and then attempt to adjust to what others competence by moving away from their anchored rating (Kruger 1999). This explanation proposes that when people are asked questions where absolute skills tend to be high, e.g. reading, speaking English, using a mouse, riding a bicycle…, people think to themselves "I'm really good at using a mouse, I can point and click while listening to music", "I can ride for five minutes without falling," etc., without considering in any detail that the population at large can do the same task just as well. People then attempt to insufficiently adjust away from the anchor they considered (viz. how good they are in absolute terms) by trying to take into account how good others are at the task. (In parallel, the below average effect appears when participants are asked how they fare versus the population at large on tasks where absolute ability tends to be low [e.g. software programming, playing chess, riding a unicycle], in which case people tend to overestimate how poorly they perform the task]).
That's the nuts and bolts of the anchoring and adjustment explanation's extension to the above average effect. Since I have been arguing at length that the Spinozan theory explains the numeric anchoring and adjustment heuristic, it seems like it would be natural for me to argue that the theory can also explain the above average effect. Moreover, the above average effect is exacerbated by cognitive load, thus displaying exactly the type of breakdown the Spinozan normally exploits. So it would appear that the above average effect is ripe for the Spinozan's picking. But on further reflection I'm not sure how to interpret the above average bias. There's something about the explanation that doesn't sit right with me. Consider the following question: must people actually *reject* their original anchor in the above average effect experiments? For it to be a Spinozan style explanation, we need evidence that the participants believed that the anchor was the answer to the question. In such a case, people will have to then adjust away from the anchor by rejecting it. It is this rejecting stage that the load exploits. But in the above average effect cases, we don't have enough evidence one way or the other to say whether people ever reject the anchor. Consider the following two cases: suppose we ask someone whether she's an above average walker. If she thinks I'M PRETTY DAMN GOOD AT WALKING, and then thinks about the general populations' skill level before making a comparative judgment, then she needn't reject the anchor, because the anchor is actually an accurate belief (I suppose that if you can walk without constantly tripping, then you're pretty good walking, at least in absolute terms). In this case the load might exacerbate the effect merely by making the integration of different information more difficult (with the different information being one's beliefs about how well others walk). On the other hand, if she initially thinks I'M PRETTY DAMN GOOD AT WALKING SO I MUST BE ABOVE AVERAGE and then adjusts to other's walking competences, then she will have to reject her earlier belief, in which case load can shut down this rejecting process. In the latter case, but not the former is rejection evident. If we knew that all cases worked as the latter case, then we'd have some stronger evidence in favor of the Spinozan view. However, as of now, I have no particular reason to prefer one story to another, and I suspect that some people make the comparative judgments before considering others' competences (like on the second story) and others don't. Additionally, part of the motivation for the Spinozan story in the numerical anchoring and adjustment paradigm was that the anchors are so arbitrary and implausible that the only way one should believe the anchors are the answer is because of some brute architectural process. However, here one could sensibly reason their way to believing the anchors, because, at least in the first scenario, the belief that serves as the anchor is *true*. Because of all this murkiness I don't know

3.4 Yea-Saying, Nay-Saying and the Need for Cognition

As we saw in section 2.5.4, yea-sayers and nay-sayers provide a fruitful testing ground for the Spinozan hypothesis. Yet the origins of yea-saying and nay-saying weren't discussed there. What is it about yea-sayers and nay-sayers that make them apt to fall into one group or the other? It couldn't be (e.g.) their tolerance for dissonance because both groups end up giving inconsistent answers to the experimenter's probes. Thus both seem to have a high-tolerance to dissonance. It couldn't be that one group really wants to make the experimenter happy while the other doesn't, because both groups confound experimenters; if they just wanted to make the experimenters happy they would be appropriate responders.

Perhaps we can make progress on this question by hypothesizing that being a nay-sayer is hard work. For us pessimistic nay-sayers, we have to overcome our tendency to accept everything we think. But being a yea-sayer is easy. All one has to do is just roll with whatever it is one thinks. There is some reason to think that this analysis is on the right track and to see why let's peruse the connection between yea-saying, nay-saying, and the 'need for cognition.' It has been noted as somewhat of a curiosity that nay-sayers are apt to have high scores on the 'Need for Cognition' scale, and yea-sayers are apt have low scores (Cacioppo and Petty 1982).[58] This is the exact correlation that one would predict if we supposed people had a Spinozan mind. The reasoning proceeds as follows: those who acquiesce more do so because they are disposed to refrain from expending mental energy, so they end up believing whatever they token and they rarely check and reject these beliefs. Those who nay-say do so

what to say about whether the extensions works for the above average effect, thus my plea for neutrality on these issues. Although the other extensions listed above fair a bit better in my mind, the evidence here is scarce enough to warrant further neutrality. (Specifically, my worries about the other effects are that they can be explained by the vividness of the anchor in a way that is closed off to the theorist trying to explain the numerical version of the effect).

[58] The short form of the scale gives participants eighteen statements to rate like "I prefer to think about short, daily projects to long term ones" (Cacioppo and Petty, ibid.).

because they are disposed to engage in strenuous mental exercise and thus are willing to expend more mental energy, making them more apt to reject their extant beliefs. Since rejecting propositions is an effortful mental endeavor, those who are more apt to reject propositions should also be more apt to engage in effortful cognition; likewise, since accepting a proposition is an effortless endeavor, those who are apt to yea-say should also be disposed to not want to engage in effortful cognition.[59] Thus, the Spinozan theory can help explain the connection between yea-saying, nay-saying, and need for cognition.

3.5 Source Monitoring Errors, Recovered Memories, and Stereotype Activations

'Source monitoring' is the name for the phenomena whereby someone remembers the source of a memory (i.e., when and where a memory was created). Confusions between real and imagined memories are generally seen as failures of source monitoring (Schacter et al. 1997). Source monitoring is particularly important in cases of recovered memories of abuse. In these cases, a therapist cues patients and prods them to remember (or sometimes 'remember') traumatic experiences that they have forgotten (or 'forgotten'). Although it's unclear whether any of these cases of recovered traumatic memory are veridical, it is clear that many of the supposed cases of recovered traumatic memory are not veridical. In these cases, the patients create, rather than recall, the event. The patient comes to 'recall' the event only after a therapist's suggestion and the patient fails to appropriately monitor the source of the memory.

The Spinozan theory can help to partially explain non-veridical recovered memories. In the studies reported in 2.5.1, participants often forgot what can be fruitfully be interpreted as the source of the sentences they read. In those studies participants would, when under

[59] Unrelated bet: I bet that philosophers score higher on the NFC scale than the average person.

load, forget the source tag of 'True' or 'False.' Importantly, when they forgot the tag, they were much more apt to remember the sentences as true than as false. The Spinozan hypothesis explains this by (surprise!) positing that subjects automatically believe what they perceive and then, when under load, participants lose the ability to use the source information (i.e., that the proposition was marked as false) to reject the proposition. The Spinozan theory can explain the recovered memories phenomenon in a parallel fashion. In the typical recovered memories situation, the source monitoring error (i.e., when people can't remember if the event happened or just is an experimental suggestion) occurs because people automatically believe whatever they entertain and, since they are under load (like one normally is in a tense therapeutic session)[60], they don't have the requisite cognitive energy to reject the propositions they entertain.

Additionally, recovered memory scenarios have a very robust stereotype. The patient generally has some negative feelings built up generally towards an older male figure, like a father, uncle, or priest. These figures also have quite stereotypical traits that are easily conjured up (for the sake of decorum I won't describe the stereotype in any further detail). The combination of stereotype activation and cognitive load, in addition to the Spinozan mind, make for a volatile situation. In a study on stereotypes and source monitoring Sherman and Bessenoff (1999) found that when under cognitive load, participants are apt to default to judgments that fit a stereotype even if they were just shown that the stereotype doesn't hold for the case at hand. For example, participants were given statements about a particular skinhead, Bob. Ten of these statements portrayed Bob as a friendly person (e.g. 'Bob gave a

[60] Merely worrying about how the therapist perceives you should be enough to create the requisite load. Of course, additional load will be brought on by regulating one's behavior (which surely is apt to occur in some of these situations).

stranger a quarter to make a phone call') and ten portrayed Bob as an unfriendly person (e.g. 'Bob shoved his way to a center seat in a movie theater'). Before participants read the statements, they were told whether the statements were true or false. During this phase of the experiment half the participants underwent cognitive load (they had to memorize an eight digit number) and half didn't. During the second phase of the experiment participants were tested on the previous statements they read. The participants who read statements under load were much more likely to remember behavior which fit the stereotype as true (i.e. they were more likely to remember Bob as unfriendly because he's a skinhead) than were the unloaded participants. Thus, participants were much more likely to make source monitoring errors when a) under load and b) when the proposition they were considering matched their antecedent stereotype.

Now imagine that you are a patient whose therapist suggests that your fear of penguins was caused by a local clergyman. In the increasingly intense therapy session, you are apt to be self-regulating your behavior (you don't want your therapist to think you're crazy, do you?), thus you're apt to be under load. When your therapist suggests that your priest might have molested you, you are more apt to make the source monitoring error because pedophile priests fit a common stereotype. The stereotyped attribution then feels right and, as the 'affect heuristic' dictates (Ramerick 2002)[61], we are apt to believe that what feels right is a guide to the truth. Now just add the thesis that you are apt to believe what you think and you get a recipe for false recovered memories. The cognitive load makes source

[61] The affect heuristic has been posited as an automatic judgment heuristic (thus different from deliberate choice heuristics, like elimination by aspects, Tversky 1972), one that guides decision making by the 'feel' of the options. One can, if one would like to, see automatic heuristics as 'system 1' processes and deliberate heuristics as 'system 2' heuristics. Not myself being a big fan of 'system 1/system 2' talk, I'd prefer not to think about them in those terms.

monitoring difficult, so you default to the stereotype and consider that the priest actually is a pedophile. Then, since you're under load you don't have the energy to check this proposition against other memories you have, so you just end up believing it without further investigation. The Spinozan hypothesis explains an important link in the chain, namely why you believe what you are merely considering.

The interaction between stereotypes and cognitive load is exacerbated because episodic recollection is more demanding and effortful than semantic recollection (Tulving 1983). When patients are asked to recall traumatic memories they are being asked to recall episodic memories. However, when load is induced, this recall is quite difficult.[62] Semantic recollection, on the other hand, is much less effortful and can occur under load. Thus, when people are put under load they are apt to resort to their stereotypes which are stored in semantic memory (e.g., skinheads are bad people) while lacking access to their actual episodic memories (and thus not making it feel that weird that they are creating, rather than recalling the memory).[63] People then believe the stereotype to hold in this case merely because they considered it while undergoing cognitive load.[64]

3.6 The Efficacy of Self-Affirmation and the Problems of Stereotype Fulfillment

[62] Try recalling an episodic memory (what the first half hour of your last birthday dinner like?) while reading this essay. Now try to recall something from semantic memory while reading the text (ask yourself what the capitol of New York is). The difference in comprehension of the text should be clear.

[63] Episodic memory in particular has its own particular troubles with recall. As Prinz writes, "A similar conclusion can be drawn about introspective access to episodic memories. Memory is a constructive process, which is prone to error (e.g., Roediger and McDermott, 1995; Schachter et al., 1998). If we report a memory, care must be taken to confirm that the memory is accurate, and introspection can easily mislead because false memories can be experientially indistinguishable from real ones. It does not follow, however, that we are inaccurate at introspecting the images that come to mind when we have what we take to be a memory experience. In the case of false memories, we may have perfectly good access to an imagined scenario and relatively bad access to the knowledge of whether that scenario actually took place." (Prinz 2004, p. 54).

[64] Recently, Bryce Huebner (2009) hypothesized that the Spinozan theory can explain stereotype activation in a wide range of cases.

The efficacy of self-affirmation is really quite puzzling. 'Self-affirmation' can be used to denote a wide range of effects. Although I think the Spinozan theory has something to offer most of the effects I will constrain my discussion to just focusing on the aspect of self affirmation that involves the process whereby one tells oneself a positive sentiment and then the sentiment seems to take hold in one's cognition. To put it crudely the therapeutic advice of self-affirmation, viz. saying what you want to believe over and over again, actually works on most people (Steele 1988; see footnote 65 for the exceptions). It is strange to think that just saying over and over again 'I'm a good, smart, likeable person' would make a difference to one's beliefs about one's goodness, intelligence, and likeability. Likewise, reminding oneself that you fit a given stereotype can make one perform poorly. For example, if one is reminded that she is a woman before taking the math section of the GRE she will perform significantly worse than if she is asked a question about her gender after she's taken the test (see Sherman and Cohen 2006 for a an overview).

These are some very strange data. Note that telling myself to not (e.g.,) fall madly in love seems to have no effect on whether I do or not; likewise for telling myself not to smoke that cigarette, eat that cookie, kick that puppy (whoops!) etc. Giving oneself commands seem to have very little effect on behavior, yet rehearsing propositions to oneself has a drastic effect. So, what's going on here?

I think that these two effects, both the positive (the efficacy of self-affirmation for forming beliefs and improving one's well-being) and negative (performing worse than one is capable of when reminded that a negative stereotype applies to oneself), can both be illuminated by the Spinozan theory. Let's consider them in turn. The positive effect is a case where people (e.g.,) tell themselves that they are good, competent people and then, over time,

they start feeling like good competent people.[65] It is difficult to suppose that people entertain the belief and then reason their way there, especially considering that the effect holds over people who have relatively low self-esteem. Presumably these folks are not necessarily apt to see themselves as good competent people. Instead we can analyze the situation as follows: when people repeat a mantra like 'I am a good competent person' they a) entertain the thought that they are good competent people, and are b) undergoing cognitive load (for they are partaking in a controlled, serial thought pattern, constantly repeating the mantra), so they disable their ability to reject the proposition they consider, thus getting the desired positive effect.

A similar explanation may apply to (e.g.,) a woman who performs worse on a math test after being reminded that she is female. Presumably, unconscious associations form a chain of activations starting from the activation of the gender construct and leading to behaviors that are stereotypical for the activated construct. Since the woman is in a mathematical testing situation, female mathematical stereotypes are apt to be tokened, chief among these the stereotype that women are bad at math. Between self-regulating one's own behavior, dealing with one's anxiety, and preparing for the ensuing exam, the stress of the testing situation imposes a fairly serious cognitive load. Thus the stereotyped belief gets activated and because of the load the woman cannot go back and reject the preposterous proposition. Since the proposition is both believed and, at that moment, active, we should not be surprised when we see its effects leak out into her performance.

[65] There are exceptions to this rule. It seems that if one starts out with super-low self-esteem and then goes through the self-affirmation process, they will come out of it feeling even worse about themselves (see Wood et al. 2009). The moral here as elsewhere is that there are very few avenues of solace for the despondent.

3.7 Negation[66]

A Spinozan theory that accepts the hitherto scarcely discussed property five makes many predictions regarding negation (as a reminder property five was: "To negate a thought is to, in part, reject it"). Two predictions in particular are germane to the ensuing discussion: the prediction that negations are difficult to process and the prediction that negations are held back in the initial processing of a sentence.

3.7.1 Explaining Why Negation is Hard

On the Spinozan theory rejections can occur only after acceptances. But it's not just the greater number of steps that makes rejection difficult; rather, it's that since starting the rejection process is optional, one has to put in effort every time one rejects a proposition. The

[66] This section, and the subsections therein, will focus on negation and not other linguistic phenomena. This is a strategic decision: I think that the explanatory work the Spinozan theory can do for negation is much more impactful than elsewhere in linguistics. However, others have taken certain linguistic data to argue in favor of the Spinozan theory. Since here I am just attempting to show the explanatory generality of the Spinozan theory I won't spend time in the main text focusing on this evidence since it falls outside the scope of my abductive argument. Yet in this (long) footnote, I'll mention some other linguistic phenomena that have been put forward as evidence for the Spinozan theory. Assume a basic psycholinguistic assumption that the complexity of thought is mirrored in complexity of language (Clark and Clark 1977 p523; n.b., I am not sure how confident I am in this assumption; consequently, I'm not sure how seriously to take some of the straws in the wind I'll now discuss). Since the Spinozan supposes that affirmations are prior to negations in processing, it might be reasonable for such a theorist to suppose that affirmations are also conceptually prior to negations. And there seems to be some supporting evidence for this claim: in general, unmarked words are the affirmative, marked the negative. Unmarked words tend to have fewer morphemes than their marked counterparts (e.g., 'happy' vs. 'unhappy'). Additionally, unmarked words may be used neutrally as opposed to marked words: 'How unhappy are you?' implies that you aren't happy at all whereas 'How happy are you?' implies next to nothing about the speaker's take on your mental state. Likewise, 'How tall are you?' at most presupposes that the speaker thinks you have some height, whereas 'How short are you?' implies that the speaker thinks you are short. These facts hold because (e.g.) 'happy' and 'unhappy' are degrees of happiness, not degrees of unhappiness (compare 'long' and 'short'—they're degrees of length not of shortness.) Marked and unmarked words are particular degrees along a dimension but only unmarked words can denote the dimension itself (Gilbert 1991). Prima facie, it seems that unmarked words are more basic than marked words, which is what we may expect if we thought that affirmation was prior to negation. Unmarked terms also denote concepts that are more conceptually basic than their counterparts: we ask if a proposition is 'acceptable' or 'unacceptable' not if it's 'rejectable' or 'unrejectable'; we ask whether something is 'true' or 'untrue', not if it's 'false' or 'unfalse'; people hope their ideas are right, not 'unwrong' and we speak of 'belief' and 'disbelief' not 'doubt' or 'disdoubt' (or I guess, 'undoubt'?). Moreover, as Horn pointed out (1989) every negative statement implies a corresponding affirmative statement, but the converse doesn't hold. Negative statements seem to be about positive statements in a way that positive statements do not seem to be about negative statements but about the world. As Clark and Clark write, a negative statement is "like an affirmative supposition and its cancellation all rolled into one" (1977).

effort needed to reject a proposition is thus greater than the effort needed to accept a proposition. Since negations are just a subset of rejections, applying a negation should also be an effortful, and thus difficult, task. This is a theoretical coup for the Spinozan because practically anywhere one looks, one can find data that negation is hard to process.

For example, adding negations to a sentence exponentially increases the difficulty in understanding the sentence with each additional negation. One doesn't need much data to see the point: it's easier to understand 'Jane kicked the ball' than it is to understand 'Jane didn't kick the ball,' which is much easier still than 'Jane didn't not kick the ball,' which in turn is easier than 'It is not the case that Jane didn't not kick the ball,' etc.

Negations also wreak havoc when they're used as a search criterion: people sort much faster and more accurately when they're asked to use a criterion that is positively formulated rather than negatively formulated (Wason 1972). Say, I wanted to use NOT-X as a criterion for a search. The Spinozan posits that every time I token NOT-X I token X, then negate it. Since the Spinozan (of my flavor) considers negations to be a subset of rejections and consider rejections to be effortful, then every negation should add increased strain to one's cognitive system. Thus, in searches (as elsewhere), I would use less mental energy if I searched for X's than if I searched for NOT-X's, for using the NOT-X criterion will always be more effortful than the positive criterion. In sum, the Spinozan expects people to flip negative criteria into their equivalent positive formulations whenever possible.

And such flipping seems to be the rule rather than the exception. For example, when people are given negative propositions that are to be used to coordinate action, they generally turn them into the corresponding equivalent positive statement before they act. Suppose I give you a deck of cards and ask you to sort the deck into two piles. You will do much better

(i.e. sort faster and more accurately) when being asked to use a criterion that is positively formulated rather than negatively formulated. Thus, people are much quicker and more accurate when being asked to sort out the spades and hearts than when asked to sort the non-clubs and non-diamonds (see Wason and Johnson Laird 1972, or Fodor 1975 for a polemical review). We would expect both faster performance and fewer errors when using a criterion that involved less mental energy, and the Spinozan theory states that the processing of affirmations uses less energy than processing their negative counterparts.

Lastly, even when people aren't performing a sorting task per se, the processing of negation is more difficult than the processing of the corresponding affirmative. In tasks like statement verification (Wason 1961) or picture verification (e.g. Slobin 1966), people are much worse at processing negations than affirmatives. This should be expected if using negated mental formulae is a more active processing than using their affirmative counterpart.

3.7.2 The Psycholinguistic Processing of Negation

The second main prediction of the Spinozan theory regarding negation is more tendentious, though there is evidence that suggests that the prediction is accurate. The Spinozan predicts that negations, as a subspecies of rejections, can only be added to whole propositions and this addition should be completed only after the proposition is formed. That is, the Spinozan theory predicts that in sentence comprehension people should process negative statements initially as affirmatives, processing the negation only secondarily. This prediction was verified in Hasson and Glucksberg (2006). There participants received affirmative and negative assertions and were then asked to perform a lexical decision task. For example, participants were asked to read sentences like 'The kindergarten is/isn't a zoo'

and ‘lawyers are/aren’t sharks.’ All of the statements participants read were metaphors, as to not allow for regular semantic priming effects to affect their data.[67]

After reading the statements the participants would see a string of letters on a screen and they were asked to assess whether the letter string spelled an English word or not. The experimenters varied the delay intervals between the metaphors and the lexical decision task and then looked at the participants’ response times. Responses to affirmative-related targets were significantly faster than negative-related targets. Furthermore, the response latencies showed that *both* affirmative and negative sentences facilitated affirmative-related primes. However, the negative related primes were not facilitated in the affirmative sentences. For example, the negative sentence ‘surgeons aren’t butchers’ equally primed the affirmative-related prime ‘clumsy,’ as it did the negative-related ‘precise’, whereas the affirmative sentence ‘surgeons are butchers’ primed ‘clumsy’ but did not prime ‘precise.’ The negative-related prime ‘precise’ only arose in the negative context, whereas the positive related prime arose in both contexts. This evidence shows the type of asymmetry the Spinozan hypothesis predicts and lends strong evidence to the view that negations are processed by first processing the corresponding affirmation.

[67] Metaphors were used so as to get around certain experimental confounds based on mere semantic priming. Imagine we gave people the sentence ‘The Man wasn’t laughing’ and then found out that HAPPY was primed. We would be unsure as to why HAPPY was primed: was it primed because negations are held back in linguistic processing so that the person first tokens ‘The man was laughing’ and then adds a negation? Or was HAPPY primed merely because its association to LAUGHING? By using metaphorical sentences, one can get at the actual processing of negation while controlling for mere semantic priming effects. This is because metaphors prime terms that aren’t primed by the metaphor’s topic or vehicle. Note that if the sentence ‘surgeons aren’t butchers’ primes ‘clumsy’ it can’t be because of the lexical associations between ‘butcher’ and ‘clumsy’ (butchers aren’t stereotypically clumsy) nor between ‘surgeons’ and ‘clumsy’ (thankfully, surgeons aren’t stereotypically clumsy either). Thus, in such a case it is reasonable to suppose that the priming occurs because of the psychological processing of negation, with negations being held back from the initial processing of the sentence.

The preceding evidence shows that people process affirmatives quicker than, and prior to, their negative counterparts. When processing a sentence, the negation is held back from the initial processing and appears online only after the initial processing happens. Negations are not initially integrated in the construction of sentence meaning. Hasson and Glucksberg's study gives us a glimpse of the actual time it takes negations to be processed. They conclude that negation doesn't take hold in processing until between 500 and 1000 ms after the negative sentence has been read, which is an enormous amount of time in linguistic processing. To illustrate, Hasson and Glucksberg non-metaphorically assert "we found that terms related to the affirmative meaning of the metaphor were accessible immediately after reading the affirmative metaphors, indicating that the affirmative meaning was arrived at immediately" (p1027; for other work showing that affirmatives are processed immediately see Blasko and Connine 1993). The Spinozan view (but not the Cartesian) predicts this startling psycholinguistic data.

3.7.3 Negation and Innuendo Effects: Once It's Out There, It's Out There

The evidence reported in 3.8.2 leads us to conclude that in order to process negative sentences we must first process the affirmative sentence, initially withholding the negation from semantic processing. It would seem natural for such a view to predict that if our processing is broken down, we should only end up with having processed the corresponding affirmative sentence. Taking this consideration seriously, as the psycholinguistic evidence forces us to, may allow us to explain the psychological phenomena known as 'innuendo effects.'[68] For example, in Wegner et al. (1981), participants read headlines about fictitious

[68] For current purposes let an innuendo be operationalized as "a statement about something combined with a qualifier about the statement" (Wegner 1984, p1). According to this criterion, a statement like 'Bryce is a pinko' is not an innuendo, whereas 'Bryce is not a pinko' is one.

characters and were then queried on the impression they formed about the character. When participants read a headline containing an innuendo (e.g. “Bob Talbert is not connected to the mafia”), they formed a more negative impression about the person than when they read a neutral sentence about the person (e.g. “Bob Talbert arrives in town”). Although this is a surprising datum, an even more shocking one was found: participants whose impressions were created by an innuendo formed an impression of the character that did not significantly differ in negativity as the impressions they formed that were created by straightforwardly incriminating statements![69] The innuendos were effective even when they were known to be coming from a disreputable source. In a different variation of the study, participants were told that the headline came from either a respectable source (The New York Times) or a less reputable source (The National Enquirer). The innuendos were equally effective regardless of source.[70]

Most research on innuendo brings one to a similar moral. For example, Wegner, Kerker, and Beattie (1978, cited in Wegner 1984) asked participants to read letters of recommendation for applicants applying to graduate school and subsequently rate the quality of the applicant. Some letters contained innuendos (e.g., “I don’t believe Paul was responsible for the loss of the laboratory tape recorder”), whereas other did not. The

---

[69] (!)Take a moment to think about how truly odd this scenario is. Reading “Barack Obama is connected to the mafia” does not make one form a significantly worse opinion of Obama than when one reads “Barack Obama is not connected to the mafia.” This is pretty astounding. Even more astounding, the same moral seemed to hold for positive innuendos (‘John gave an elderly man some money’ vs. ‘John did not give an elderly man some money, Beattie and Wegner 1980, reported in Wegner et al. 1981).

[70] This leads to an important practical point about propaganda. Say you run a less than wholly reputable periodical. If your headlines are statements without qualifiers (e.g. ‘Bryce Huebner eats babies’) then people will be more apt to think hard about the statement (since it’s so strikingly false) and thus think about the source and ignore the content as propaganda. However, if the same periodical just brings up the innuendo (e.g., ‘Bryce Huebner does not eat babies’, or ‘Does Bryce Huebner eat babies?’) readers will be more likely to form the (desired) negative impression in part because they are less likely to think hard about the statement.

participants who read the letter with the innuendo rated the candidate as significantly less acceptable than the candidate whose letters did not contain an innuendo.

The Spinozan theory conjectures that these innuendo effects arise because people are actually tokening the denied proposition BOB TALBERT IS CONNECTED TO THE MAFIA, and hence believing the proposition, on the way to tokening the negation (and thus believing its denial). Innuendo effects arise because of the 'transparency of denial' (Wegner et al. 1985), whereby people seem to 'see through' denials and thus process the counterpart affirmation.[71] It is possible that the way cognition evolved was first by creating minds that could believe, and only later developments allowed for the ability to negate beliefs. The current suggestion is that perhaps what held for evolution also holds for cognitive and linguistic processing, for it seems that in both arenas affirmatives are the basic grounds for processing with negations being processed only subsequently.[72] If the preceding line of thought is on the right track, then the Spinozan theory can explain why people are so frustratingly susceptible to innuendos.[73]

3.8 Fearing Fictions

The fearing fictions phenomenon is a well-entrenched philosophical puzzle. Perceived fictions seem to play a paradoxical role in our mental lives. For example, when

---

[71] Not all denials are equally transparent. People have an easier time ignoring innuendos containing abstract statements ('Bryce is not nice') from innuendos containing concrete statements ('Bryce did not hold up a bank'). Presumably this is because once someone negates an abstract statement, there is a corresponding positive idea that can arise (e.g., Bryce is cruel), whereas there is no easy counterpart to (e.g.) not holding up a bank.

[72] The foregoing analysis leads us to a prediction: if the participants in the studies above were to be put under load, then they should show even more severe innuendo effects. That is, their ratings of the characters between the (e.g.) negative innuendo headline and straightforwardly negative headline should be near identical when someone reads the former headline while under load.

[73] In the same spirit, the theory can help explain the mechanics of propaganda transmission, a topic that will be discussed separately in chapter 5.

watching horror movies, how can one be scared by the monster on the screen while simultaneously not believing in monsters? It is often assumed that this is possible because people act 'as-if' monsters exist, willfully suspending disbelief (Walton 1978). However, given that people don't have direct volitional control over their beliefs, the *willful* suspension of disbelief seems impossible. If that's so, then how do people accomplish being afraid by a film they know to be false?[74]

According to the Spinozan model, since we automatically believe what we perceive, we are afraid of scary fictitious films because we believe what we are watching is veridical; watching the monster on screen actually yields the belief that there is a monster. Thus, we are afraid. Of course, we also believe that there are no monsters, but in this context, the immediate perceptual input makes the monster belief more salient than its negation. What prevents us from fleeing the theater is that we remind ourselves of our stronger belief that there are no monsters (sometimes by repeating the mantra), thereby changing the saliency of that belief. The Spinozan theory has the resources to diffuse the paradox of fearing fictions quite handedly. All it asks for is a notion of salience (which we'll need for our cognitive science anyway). One may object that the theory also must posit that people hold contradictory beliefs in order to explain feared fictions. And this is true: the Spinozan theory is up to its ears in positing contradictory beliefs. However, this shouldn't worry you because there is ample evidence that people do hold contradictory beliefs (for a defense of this claim see the 'alief' objection in section 4.5).

3.9 Summary

[74] There is a novel explanation of the fearing fictions phenomenon due to Gender's notion alief (Gendler 2008a and 2008b). However, I do not think that we should countenance any notion of alief (for more on my objections to aliefs see section 4.5).

In this chapter, we've surveyed a wide range of data, ranging from the venerably recalcitrant to the utterly mysterious. In each case, the Spinozan view has been able to shed light upon previously poorly understood phenomena. The generality, elegance, explanatory power, and general fecundity of the Spinozan theory give us reason to believe that the theory gives us an accurate description of our belief fixating process. The Spinozan hypothesis is not just the best going theory we have of how beliefs are acquired, but it also can explain a slew of other effects outside the arena of belief fixation.

Perhaps more impressively, in some of these instances, the Spinozan theory explains how certain 'higher-level' phenomena observed in social psychology arise due to architectural constraints that are 'one level down,' so to speak. Three of the most robust, well-known social psychological phenomena, the fundamental attribution error, the confirmation bias, and the anchoring and adjustment heuristic, can be understand as a natural consequence of a cognitive architecture where everything that is tokened is believed. Thus, the Spinzoan theory allows for a glimpse at the holy grail of cognitive science: understanding higher level phenomena in terms of some lower level constraints. Note that I'm not arguing that these higher level effects are *reducible*; rather, it's that a certain architectural constraint, that thinking causes believing, can be used to explain various higher level phenomena of different stripes (although the fundamental attribution error may be in the purview of social psychology [whatever exactly that means], the psycholinguistic processing of negation is a phenomenon in cognitive psychology). In sum, if the Spinozan hypothesis is true, our understanding of what appeared to be quite different psychological processes will all be explicable with one elegant theoretical posit: thinking is believing. And this is all icing on the cake, for the theory is independently motivated as the best going explanation of the data we

have on belief fixation. However, that does not mean that theory doesn't have its problems and critics. In the next chapter I will go through some of the main objections that have been offered against the Spinozan theory. Then we'll end by taking stock of the picture of the mind that emerges once the theory is taken seriously. But before the speculating begins, let's have some gladiatorial debate.

Chapter 4: Objections and Replies:

Imaginary Conversations with Real Critics

The Spinozan theory strikes many as deeply radical and unintuitive. Consequently, it is incumbent upon me to not just respond to specific criticisms, but also to try and explain why so many people find the theory to be so radical. In this chapter I'll survey a range of objections to the theory. I'm sure that there will be some objections that I'll miss, but I'll try to canvass at least the main objections put to the theory. In responding to the objections (particularly the first objection) I'll try to piece together where people's intuitive biases against the theory stem from. In the final chapter, I'll speculate a bit further about why I think the theory strikes many as obviously false. But before we get there, let me attempt to assuage the numerous philosophical and empirical objections to the view.

4.1 The Objection from Introspection

Let's start with the objection that arises most frequently. The objection starts by someone attempting to consider some fantastically odd proposition, like *dogs are made out of paper*. The objector then proclaims that after some quick consideration she is sure that she does not believe that dogs are made out of paper. Since the Spinozan theory says otherwise, she concludes that the Spinozan theory is false. Sometimes an incredulous stare is added for good measure.

The intuition behind this type of argument is fairly robust. In general, people think that they know what they believe and they know it straightforwardly through introspection.

This intuition presupposes that beliefs are the types of things that are consciously accessible through introspection. However, I, following many self-respecting philosophers and psychologists, do not think that beliefs are in general accessible through introspection (e.g., Bem 1970; Lycan 1986, forthcoming; Williamson 2000).[75] For example, Bill Lycan writes,

> I doubt that one can introspect one's beliefs or other merely dispositional states. If I want to find out what I believe on such-and-such a topic, I ask myself the question and find myself making a judgment. The judgment probably—not necessarily—manifests my existing belief and so reveals it to me, but that process does not count as my introspecting the belief itself (p. 6).[76]

Along delightfully similar lines, when discussing Daryl Bem's work on belief Joel Cooper writes,

> We do not always have insight into our own attitudes and beliefs, especially when they are not very strong or salient…When asked about our opinion toward most political issues or attitude objects we engage the very same process to infer our attitudes as we use to infer the attitudes of others. We look at our behavior, analyze the environmental stimuli, and make a logical inference about our attitudes. (Cooper 2007, p.37).

[75] Similar suggestions can also be found in Evans (1982), Crane (2001), and Prinz (2004). For example, Evans thinks that when we think we are introspecting our stock of beliefs we are actually just figuring out what it would be rational to believe, a view that's not all that dissimilar from what I'll be arguing for. Prinz is a bit cagier on exactly how accurate he thinks our introspective capacities (for what he calls "attitude access") can be. But he seems to agree with the main point I'll be arguing for. On our access to beliefs Prinz writes, "We gain conscious access to beliefs by figuring out what we are disposed to say. It is difficult to speculate about whether this process is accurate, but it certainly seems to leave room for error. If we do not experience the process by which a belief is summoned, nor the full mental representation of the belief itself, we may be getting incomplete, biased, or otherwise degenerate information in consciousness…With belief, we often have but one thing in consciousness: the belief report without the belief itself. Room for error seems greater then. I think the fact that we tend to discover our beliefs indirectly by discovering what we are going to say may help to explain why the concept of belief is more difficult to acquire than the concept of desire or the concept of perception" (Prinz 2004, p.55)

[76] This line of thought has been alive in Lycan's work for some time: "It is an interesting question whether we can ever introspect beliefs. On both phenomenological and theoretical grounds I doubt that too; what we introspect, in the way of cognitive items, are judgments, and we infer our knowledge of our beliefs from these" (1986, p 64).

If I ask you whether you like pinto beans, you may immediately know the answer (perhaps you have a pinto-based diet), but more likely you recalled your history with pintos. Perhaps you ordered them last week, so you infer you must like them or else you wouldn't have ordered them. We generally infer what we believe by examining our past behavior (even if such an examination is reflexive, unconscious, and instantaneous). Of course, in the paradigm instances of belief, the belief has been made salient to us so often that we needn't engage in any elaborate inferential process: if I ask you whether you love your spouse, you generally know what the response is; if I ask you whether you believe that 2+2=4, you can quickly respond because it is a question you have frequently answered.

The intuition that we have the ability to introspect our beliefs is a cognitive illusion caused by the paradigmatic cases of belief. When we are asked what we believe about a topic that is strongly affectively valenced the answer arises instantaneously. Yet, the vast majority of the beliefs we hold are not strongly valenced like our belief that genocide is wrong. Rather, the vast majority of our beliefs are more like our belief in the tastiness of pinto beans. It is the salience of paradigm cases that leads us to infer that all cases of belief are like the cases of strongly valenced belief. Once one spots how the salient cases differ from the majority of cases, the intuition pushing against the Spinozan view should be tempered.

Moreover, one can read the last thirty years of psychology as one long story about how opaque our minds are to ourselves. Philosophers have overplayed how much we can introspect because they often parochially focus on the contents of thought as opposed to mental processes. Mental contents are sometimes available for report (of course, sometimes they aren't too). However, our thought processes are almost never available to report (see, e.g., Nisbett and Wilson 1977). If they were available, then (e.g.,) the cottage industry of

priming studies wouldn't be thriving and cognitive illusions wouldn't be so surprising. Not only are our mental processes unavailable for report, but we are even apt to misreport our emotional states (Dutton and Aron 1974). One would think that the propositional attitudes would be more like emotional states and mental processes than contents. Even our metaphors for the attitudes (e.g., the 'belief *box*') are more like processes than contents—after all, the belief box is supposed to be the 'place' you put certain contents. Beliefs, like propositional attitudes in general, attach to contents. A belief is a content with a certain functional role and having a functional role is playing a part in our mental economy. Accordingly, functional roles should strike us as quite similar to mental processes (both are operations on contents, not contents themselves) and as such shouldn't be introspectable. In sum, I'd like to offer this as an observation: for some time, the cognitive science community has been moving in the direction of thinking that our access to mental operations is, at best, quite fallible. However, we intuitively think that our access to our mental states is quite good, and sometimes totally diaphanous (another, different, though not unrelated Cartesian dogma). This is because our access to our mental contents feels strikingly clear.[77] But propositional attitudes, like mental processes, are not contents, but rather mental relations to contents. Although we have some (though certainly not as great as advertised) evidence that we can introspect our contents, there is no evidence that we can introspect our *relations* to contents, which is what beliefs are. Moreover, since we know we are quite incompetent at introspecting other cognitive relations[78] we should be Very Skeptical that we can introspect our propositional attitudes.

[77] This is not meant to mean that we have perfect access to our representations. Everything I've said here is compatible with the thesis that what we experience is the outside world through our representations—I have no truck with that type of diaphonousness.

[78] One can interpret mental processes as functions from certain inputs to certain outputs, which makes them relations in a similar sense to the propositional attitudes.

Since we can't introspect our beliefs, I am unmoved by an objection that crucially relies on introspection. Rather, we find out what we believe by simulating what others would do in our position, by watching our own behavior, by inferring from past instances, etc. Our lack of introspective access to our beliefs is central to solving multiple psychological puzzles: it's why there are so many implicit racists who make sincere avowals of their egalitarian beliefs; it's why people fall in love when traveling other continents because they mistake fear for lust; it's why writing a counterattitudinal essay will sway what we report our beliefs to be... Many beliefs that we think we don't harbor, we do, only we won't be able to figure that out merely through introspection—that's why we have clever psychologists.

But my critic may persist: "But what about the belief that DOGS ARE MADE OUT OF PAPER? I just tokened that thought and I swear I don't believe it. You say I do. What evidence have you for this?" Well, to my critic I should first note that, a ridiculous belief such as DOGS ARE MADE OUT OF PAPER isn't a belief that's going to eventuate in much behavior, certainly not in the nanosecond after contemplating it and before telling me that you don't harbor the belief. Since this belief has such a low chance in causing any behavior, you couldn't come to find out that you harbor this belief even if you were excellent at reading your beliefs off of your behavior.[79] If you considered a more sensible though still outlandish proposition, such as *all dogs carry deadly viruses*, you would probably also claim to not believe it after consideration. But for all that, you would probably show signs subtle signs of harboring the belief. For instance, if after considering that proposition you were

[79] One may object by saying that the belief could show up in behavior at a later time. For example, if you believed that dogs were made out of paper then why would you ever give your dog a bath? However, the Spinozan can respond that you probably also have a much stronger belief that dogs aren't made out of paper and we'd expect stronger beliefs to win out (in most contexts) over weaker ones.

presented with dogs, you'd probably start lightly sweating, the galvanic skin response being an effect of having considered the proposition.

When you consider a ridiculous proposition, you generally attempt to falsify the correspondingly acquired belief immediately. Assuming you're not under cognitive load, you can normally do this quickly. What then becomes available for introspection is your *judgment* that dogs aren't made out of paper. From this you rightly infer that you believe that dogs aren't made out of paper (after all, the Spinozan is tied to the claim that you can't judge that x without believing that x). Thus, in many of these cases people will both believe that dogs are made out of paper and believe that dogs aren't made out of paper, but they'll only think they harbor the latter belief because they have access to the judgment that accords with that belief. However, the consideration process just serves to change the relative strengths of these beliefs. A well-informed deliberator will raise the strength of the negated belief, but will still have formed the affirmative belief.

Perhaps an example will make you feel more comfortable with idea that we don't have introspective access to our beliefs. Let me introduce you to a fictional character, Ram, an early thirty something who has always been pretty skinny and has imbibed a substantial amount of beer in his day. Ram is a skinny guy who was worried that he had a beer gut, but didn't actually have one. After Ram turned 30, he started thinking that he had one. He would walk shirtless to the shower looking down and seeing a slight bulge in his belly, from which he inferred that he had a beer gut. A few weeks later Ram made a self-deprecating joke about his beer gut to his friends, who acted astonished at the suggestion. Ram then asked his friends if he has a beer gut and his friends said that he didn't. Ram trusts his friends and believes that they are giving him a sincere response. His friends' adamant denials of the beer gut serves as

the best evidence he has; he now happily reports that he is in fact not a skinny guy with a beer gut.

However, every time Ram looks down at his stomach he *sees* a beer gut. Because he trusts his friends' words, Ram tries to discount these perceptions. For example, if Ram is asked if he has a beer gut, he asserts that he doesn't have a beer gut.[80] One might thus reasonably suspect that Ram doesn't believe that he has a beer gut. Yet if you want to predict the majority of Ram's behavior your best bet is to believe that Ram believes that he has a beer gut. When Ram walks by a mirror, he's apt to turn sideways so see if he has a bulge; when Ram walks to the shower, he still looks down and gets a spike of anxiety; when Ram approaches the buffet table he thinks twice about the au poivre sauce; when Ram sees a beer commercial, he winces; when Ram goes clothes shopping he opts for the baggy shirts as opposed to the more form-fitting ones. Yet Ram sincerely reports that he believe that he doesn't have a beer gut. So, what's going on with Ram?

One important datum in explaining this situation is realizing that Ram looks down and sees a beer gut much more frequently than he hears that he doesn't have a beer gut. Ram looks down a sees the beer gut every day, whereas Ram's friends interventions happen quite infrequently. Although Ram discounts his beer gut perceptions as optical illusions and he trusts his friends' reports that he's beer gutless, he still acts as if he has a beer gut. The Spinozan theory proposes that Ram acts as if he believes he has a beer gut because he *believes* that he has a beer gut. Since Ram continually perceives that he has a beer gut, he's continually tokening the thought that he has a beer gut, and this continual tokening is sufficient for believing that he has a beer gut. Moreover, the relative strength of beliefs

[80] As one might suspect, a consequence of the view I'm offering is that it severs the tie between assertion and belief. However, the Spinozan can still hold a tight connection between assertion and judgment.

(between say, believing one does have a beer gut versus believing one doesn't) is in part a function of how often the particular belief is tokened.[81] Since Ram tokens the belief that he does have a beer gut more often than the belief that he doesn't, he believes that he does more strongly (and hence you see it in his behavior more clearly).[82] He thinks he believes that he doesn't have a beer gut because when he's discussed the issue in the past he has come to the sensible conclusion that he doesn't. However, since he can't introspect his beliefs, he only reports the belief that seems most reasonable, which is his judgment that he does not have a beer gut.

Ram's case is by no means unique. The moral of it echoes the point made in the Lycan quote above: we need to make a distinction between belief reports, which are judgments, and beliefs. What we can introspect are not the latter, but are the former. Beliefs are unconscious propositional attitudes. In contrast, belief *reports* are a species of judgments and they're judgments that can be affected by all sorts of pragmatic factors. The beliefs that we report having are beliefs that on reflection we catalog as normatively respectable. In essence, the beliefs that we report are the beliefs that we *endorse* and we generally are wont to only endorse normatively justifiable propositions. Consequently, what we endorse is affected by a slew of heterogeneous factors like social pressure, anxiety, face-saving techniques, etc. We endorse propositions that seem reasonable to us, and when we are 'introspecting our beliefs' we are generally just reasoning about what seems rational to believe, not searching our actual stock of beliefs. What we end up sincerely reporting as

[81] To reiterate an earlier point, this is why, in part, the therapeutic advice of self-affirmation theory (saying what you want to believe over and over again) actually works (Steele 1988).

[82] Additionally, every time Ram token the negated thought *I don't have a beer gut*, he tokens *I have a beer gut*, which raises the strength of the affirmative belief.

beliefs may have little in common with what we actually believe. What we endorse is a social matter, but what we believe is a brute architectural matter: we believe what we think, even if we think many things that we would never want to publicly endorse.

One might object: "beliefs are the types of things that play a role in practical reason. How could beliefs play these roles if they are never conscious? Either your 'beliefs' don't play these roles and so aren't beliefs or they do play this role and so are available for conscious introspection." However, this line of thought is misguided. In effect, the Spinozan theory accuses the folk view of behavior of making too few distinctions. The Spinozan sees something akin to practical reasoning occurring on two levels: one at the conscious level and one at the unconscious level. At the conscious level judgments, not beliefs, play a critical role; on the unconscious level beliefs take center stage. Thus, the Spinozan can allow that beliefs still play the same role that they always did, they are just not accessible in ways we might have pretheoretically thought they were.

My critic may not be completely pacified by the above arguments. The proponent of introspection may think that I have given cognitive phenomenology a short shrift. Before I leave the topic of introspection, let's take a short spin down the road of cognitive phenomenology to tie up one last loose end.

4.1.1: Cognitive Phenomenology

My critic may complain that there is a distinctive cognitive phenomenology that attaches to beliefs. She may continue that one couldn't experience such phenomenology without having introspective access to her beliefs, thus concluding that since she experiences such phenomenology, she must have such access. Since she has such access the argument from introspection stands and she thus has the final word on what she believes.

In response I'd like to point out that one needn't be J. B. Watson in order to reject this line of thought.[83] I do not have to deny that there is some phenomenology that appears to come along with belief, but I do deny that appearances are a decent guide to reality here. Suppose I believe that tables are scary. There might be some phenomenology that attaches to the thought TABLES ARE SCARY. However, there is precisely no evidence that the phenomenology stems from the propositional attitude and not the content of that attitude. Moreover, we know that contents feel like something, so ascribing phenomenal qualities to the attitude buys us exactly no explanatory ground. My suggestion is that the 'phenomenology of belief' comes from the contents of belief and not from believing; in other words, I'm suggesting that, strictly speaking, there is no phenomenology of belief.

My response here echoes Lormand's response to Goldman's claim (1993) that beliefs have an experiential dimension. Since I agree with this sentiment whole heartedly, I quote at length. Lormand writes,

> One's standing belief *that snow is white* may cause one to think *that snow is white*, by causing one to form an auditory image of quickly saying the words 'Snow is white' (or 'I believe snow is white')…. At least normally, if there is anything it's like for me to have a conscious belief that snow is white, it is exhausted by what it's like for me to have such verbal representations, together with nonverbal imaginings, e.g., of a white expanse of snow, and perhaps visual imaginings of words. The important point is that the propositional attitudes are *distinct* from such…[phenomenally] conscious imagistic representations…. Excluding what it's like to have [the] accompanying…[imagistic] states, however, typically there seems to be nothing left that it's like for one to have a conscious belief that snow is white. (Lormand 1996, p. 246-47, taken from Lycan forthcoming, p 3.)

[83] Watson famously rejected Titchener's Structuralist school of psychology on the reasonable grounds that his introspective methods weren't replicable (Watson 1913; though Watson wasn't the only one to do so—see e.g., Dewey 1918, Dunlap 1912 for two more or less random examples. [I can't help but mention the similarities between Dunlap's view and Dretske's views—worth seeing first hand.] Of course Watson also rejected Titchener's work on less reasonable metaphysical grounds, but that is not of import here. For some interesting back and forth between them see Larson and Sullivan 2006).

The moral, to repeat, is that when we think we are introspecting an attitude we are instead introspecting (if anything) the content contained in the attitude. To the theorist who pounds the table insisting that he can indeed tell if what he's introspecting is a content or an attitude, I return to Watson's criticism of Titchener. Watson's most reasonable criticism was that introspective psychology had no adequate controls and as such was just not replicable. People could just run around (sincerely) claiming that they introspected whatever they felt like they introspected, which is all fine and good, except if you're trying to make a theory that covers more than just yourself. I've heard Uriah Kriegel unpersuasively argue against this line of thought by saying that cognitive phenomenology is just manifest in his experience(!). When I remarked to him that the Titchener and his gang of Structuralists said the same thing right before their school died he remarked that different people can differ and that he was just trying to understand his experience.[84] Which is good for him, but I don't particularly care about his experience, nor my experience, per se. Instead I care about the human experience, about psychological facts that hold across different people. Thus, if you are the type of person who just has certain things 'manifest' to yourself that others don't have 'manifest' to themselves, that's all fine and dandy, but it's not the project of cognitive science nor philosophy to explain these things. That's why God created poetry and therapists (which are two things that I think are wonderful, but are [sadly] not part of the current endeavor).[85]

[84] Along similar lines Lycan cites Horgan and Tienson. Lycan writes, "They [Horgan and Tienson] go on to announce, 'Virtually everything we have been saying is really just attentive phenomenological description, just saying what the what-it's-like of experience *is* like. It is just a matter of introspectively attending to the phenomenal character of one's own experience.' Please." (Lycan, forthcoming, p. 45). I am not sure I've ever agreed more with an argument than Lycan's one word argument here.

[85] Keeping the same theme going, Pitt (2004) has claimed that his introspection shows that propositional attitudes have a distinctive shared phenomenology. Unsurprisingly, using the same method Georgalis (2006) has claimed that the attitudes share no distinctive uniform phenomenology "all instances of believings do not

To sum up, rejecting the Spinozan view based on one's confident introspective assertions is to partake in both bad philosophy and bad history. One needn't be a behaviorist in order to see that very bad consequences arise when one takes introspective data as indisputable. I'll end this section with one fun historical example. Since it's too easy to pick on the early Introspectionists (and at least they had some positive impacts on the growth of psychology by attempting to empirically investigate experience) let's turn our focus to what relying on introspection did to the great Gestalt psychologist Kohler (who was no fan of Introspectionist psychology in general, see Kohler 1929). When Kohler was closing his investigations into the auditory sensation he notes the compelling similarity between feeling high tones as bright and low tones as dark. He then noted that others have introspected a similar impulse to describe some tactile sensations as bright and others as dark. This causes him to infer that sight, hearing, and touch must share a preponderance of overlapping cortical areas, for why else would introspection show them to have similar qualities? He concludes that there should be no dedicated neural circuitry for any of the three senses.[86] Kohler made many discoveries, but I suppose this doesn't count as one of them. Here, as elsewhere, an overreliance on introspection can easily lead to egg on one's face.

4.2 The Gullibility Heuristic Objection

seem to me to share some uniformly identifiable phenomenal or qualitative feature… I find a similar lack of uniformity when I consider classes of other occurrent attitudinal states" (p. 72, cited in Lycan forthcoming, p. 42). I believe the right response to these debates is the one that Watson would give: introspection is just too loose a branch to be relied upon without further non-introspective evidence. (For what it's worth even the arch-Introspectionist Wundt seemed to agree that introspection should only be used as a check on non-introspectively gathered evidence. Wundt wrote, ''Introspective method relies either on arbitrary observations that go astray or on a withdrawal to a lonely sitting room where it becomes lost in self-absorption',' Wundt, 1900, p. 180, translated in Blumenthal, 2001, p. 125. For more on the history of Introspectionism see section 5.2.1).

[86] This anecdote is relayed in Washburn (1922). Somewhat surprisingly, she relays this anecdote in glowing terms, commending the inference. Perhaps this shouldn't be such a surprise, since her article is title "Introspection as an Objective Method."

The evidence used in chapter 2, particularly section 2.5.1, may strike the reader as somewhat queer. The data there has been culled from situations where people encounter certain propositions while under cognitive load. One might reasonably object to using that evidence as evidence in favor of an architectural processing story that the Spinozan prefers. One might prefer is to instead posit that a pretty simple heuristic may be able to explain all of the previous data. Such a theorist might conjecture that people have a belief bias, or what I'll call "the gullibility heuristic." The gullibility heuristic would be an automatic heuristic (like the affect heuristic mentioned in section 3.6, or the representative or availability heuristic, Kahneman and Tversky 1974), which more or less amounts to a rule (whether explicitly or implicitly stated and followed) that causes one to take whatever one perceives as true. Roughly, the idea behind heuristics is the tougher the computational task, the more apt you are to default to using a heuristic. If the problem that one is dealing with is too computationally demanding (e.g., making probability judgments) then one doesn't engage in the demanding processing and instead uses a rule of thumb (like trading in representative categories for probabilistic distributions Kahneman and Tversky 1981). Figuring out what to believe is a very computationally demanding problem. It's difficult enough that it sometimes goes by a proper name: the Frame Problem. One version of the Frame Problem is the problem of figuring out which beliefs to update and which to ignore based on one's current evidence, stock of beliefs, and recent actions (Dennett 1998). Some have taken the problem to be so intractable that they see the study of belief updating (and central cognition in general) as a fruitless venture (e.g. Fodor 1983, 1987b, 2000).

But perhaps the gullibility heuristic could be used to solve the Frame Problem. Perhaps what people do is believe everything most of the time, and then later on toggle belief

strengths in different ways.[87] In short, the problem of belief updating is exactly the type of problem that is ripe for heuristic solutions, so perhaps what we should look for is not an architectural solution, but a heuristic one. Moreover, since many of the previous studies depend on getting someone to believe some stimulus that is presented exogenously (e.g., the memory asymmetry studies), perhaps those could be understood by merely positing the gullibility heuristic.

However, as Hamlet knew quite well, things aren't always as they seem. For one thing, the proponent of the gullibility heuristic finds herself in an awkward position. All of the heuristics I can think of are heuristics that are not 'plan b' reasoning strategies but are instead one's first line of defense in problem solving. For example, when people are asked to quickly figure out probabilities they often use the representative heuristic, using a prototype mental representation, instead of probabilistic calculations even when they are statisticians who are adept at making probabilistic calculations (Kahneman & Tversky 1981). Not using the representative heuristic takes effort, whereas using the heuristic is what comes naturally. Another way of putting the point: canonical heuristics are used *even when we aren't under cognitive load*. For example, the gaze heuristic (Gigerenzer 2007) is the heuristic that is used by people when they are trying to catch a fly ball. The centerfielder who attempts to catch a fly ball doesn't quickly solve a bunch of differential equations, she doesn't calculate the air resistance, the force of the swing, the velocity of the ball, the direction of wind, etc. Rather, what she does is a) starts running in the direction of the ball b) tilts her neck so that she can

[87] Of course, this way of 'solving' the Frame Problem just pushes it one step back: now the problem will arise for weighing the strength of beliefs as opposed to belief updating.

gaze at the ball, and c) tries to ensure that the angle of her neck stays constant. Following these steps, one can track down fly balls quite easily.[88]

Now the gaze heuristic is like heuristics in general in that it's apt to be active whether or not one is under load. Note that the proponent of the gullibility heuristic would be in a much different boat, for she assumes a Cartesian theory where we can entertain a proposition without believing it unless we are under load, in which case the gullibility heuristic kicks in. Thus, one who championed the gullibility heuristic would have to explain why it only showed up in cases of cognitive duress because all other known heuristics, e.g. anchoring and adjustment, representative heuristic, the gaze heuristic etc., are used almost all the time regardless of cognitive load. Although this isn't a knock-down against the heuristics viability, it should give one pause.

There are further reasons for being skeptical of the gullibility heuristic story. For one thing, it can't explain the belief perseverance effects covered in section 2.5.3. The participants in the belief perseverance studies aren't under cognitive load; they have all the time they'd like to form their own thoughts about their abilities. Furthermore, most heuristics can be overcome in certain (albeit rare) situations (e.g., Nisbett and Ross 1980). So, when participants are told that the feedback they are about to receive is false, they should withhold from forming beliefs based on that feedback (after all, what better situation could one find for shutting off the heuristic than one in which you are explicitly told that the heuristic won't hold). However, as we've seen, this is not how people behave. One would think that if the

[88] Along similar lines of thought, if you take world class centerfielders and make them sit in the bleachers and then ask them to predict where a ball will land in the outfield, they will not answer the question any better than my immigrant grandmother who couldn't tell a baseball from a football (Gigerenzer ibid.). Stationary centerfielders aren't any better at predicting where balls will land than others, but they are much better at running to the balls more quickly than others are.

gullibility heuristic were like other heuristics then it wouldn't be used in the aforementioned prebriefing situations. Thus, it seems like the belief perseverance effects are anathema to the gullibility heuristic hypothesis.

Even if we put aside the suspicious uniqueness of the gullibility heuristic and we choose to ignore the prebriefing data, there are still other unsurmountable hurdles for the heuristic story. In particular, the gullibility heuristic cannot explain why people seem to believe everything they think, even when the ideas *are self-generated* and the participants aren't under load. The gullibility heuristic also appears to be committed to some neuropsychological hypotheses that seem quite dubitable. Since both of these objections to the gullibility heuristic will take a bit of space to unpack, I will give them their own separate subsections.

4.2.1 Evidence against the Gullibility Hypothesis Part 1: Self-generated Anchors

Since the gullibility heuristic says to believe what you *perceive*, it should only range over exogenously given stimuli, so if there was evidence that the same type of effects covered in chapter 2 could hold over endogenously given stimuli, the gullibility heuristic would be obviated. Finding such data is a tricky affair, because it's always a bit unclear on how to determine from the third person perspective what one is thinking. Consequently, no one has tried to empirically test the Spinozan hypothesis using endogenous stimuli, which is understandable because it's hard to see what type of experimental design could be called upon for such a test. But one can sneak upon such data if the experimental set-up is clever enough. Happily for the Spinozan, there are some extant data that, when repurposed, shows that endogenously given stimuli are also automatically believed.

In a series of studies (Epley 2004; Epley and Gilovich 2001, 2006; Epley et al. 2004) researchers tweaked the traditional anchoring and adjustment paradigm in ways that are germane to the current discussion. Remember that in the original paradigm experimenters ask participants to figure out numerical values for some arbitrary questions, but before participants are allowed to answer the target question, the experimenter arbitrarily selects a number that serves as an anchor. Participants then answer whether the target question was higher or lower than the arbitrarily picked number. After answering, participants then state what they take to be the actual answer.

In section 3.3, I conjectured that the Spinozan theory can explain the anchoring and adjustment effects. Yet the proponent of the gullibility heuristic may respond that because the anchors are exogenously given, she too can explain the anchoring and adjustment effects. However, the gullibility heuristic explanation only holds if the anchors are exogenously given. Thus, the gullibility heuristic would be defeated if self-generated anchors can be used to create the anchoring and adjustment effect. The tweaks that Epley and co. made to the anchoring and adjustment paradigm affect exactly this question. Epley and co. coerced participants to self-generate anchors by asking them questions that they knew would produce certain clear reference points that would serve as anchors: 'When was Washington elected President?' (the anchor being 1776), 'What's the boiling point on Mount Everest?' (the anchor being 212), 'At what temperature does vodka freeze?' (the anchor being 32), 'What's the gestation period for a baby elephant (the anchor being 9)[89] etc.

[89] Over half the participants thought that the gestation period for an elephant was less than 9 months (it's around 22 months—not that I knew that without looking it up, but I was fairly confident that most people would assume it's bigger than 9 months). I'm not sure if this should make us laugh or cry. How about both?

The endogenously derived anchors caused the same type of anchoring and adjustment effect as the exogenously given anchors, and to the same degree. The results were determined as follows: say that the question is 'When was Washington elected president?' Here, the self-generated anchor is 1776 (with the actual answer being 1788).[90] Participants were asked both to give an exact response (the mean response across participants was 1779.67) and they were asked to provide a range of plausible values for the answer (the average range was between 1777 and 1784). To show that anchoring and adjustment was occurring in these cases, all we need to show is that the participants responses were skewed away from the midpoint of their plausible range and toward the self-generated anchor. After all, if one thought that the plausible range of responses for a given question was, say, 1-10, then, all else being equal, one should guess 5 as the value. If one's response was consistently skewed to one end of the range, we'd look for some reason why. In these studies participants' responses were consistently significantly skewed away from the midpoint and toward the end of the range that coincided with the self-generate anchor on each question examined.[91]

[90] One needn't just infer that the participants used the self-generated anchor as a reference point, for the participants were explicitly asked how they arrived at their answers and the participants mentioned referencing the self-generated anchor when answering the question (Epley and Gilovich 2001, 2006). The participants who didn't mention referencing the self-generated anchor were excluded from the analysis.

[91] To show that such a skew occurred a 'skew value' was determined by dividing the difference between the estimated answer and the range endpoint nearest the intended anchor by the total range of plausible values (Epley and Gilovich 2006). As a consequence of the formula, estimates that were perfectly centered within the range of values the participant deemed plausible would receive a skew rating of .5 (showing that it wasn't skewed toward either endpoint of the range). Take the George Washington example above. To find the skew rating we just calculate the difference between the mean response to the question (1779.67) and the range endpoint nearest to the self-generated anchor (since the range was 1777.29-1784.57 and the self-generated anchor was 1776, we use the lower end of the range, 1777.29) by the total range of plausible values (1777.29-1784.57). Thus our formula reads (1779.67-1777.29)/ (1784.57-1777.29) giving us a mean skew value of .33. The skew values significantly differed from .5 regardless of whether the participants themselves generated the plausible range of values (within subjects condition) or whether a separate group of participants generated the range of values (between subjects condition, Epley and Gilovich 2006).

Now we should ask ourselves a similar question that we posed in section 3.3: what could explain the anchoring adjustment effect? There, I argued that the only explanation on offer was that people actually believed the anchor to be the answer to the question because of their Spinozan minds, and then they tried to adjust away from the anchor. However, when we started considering the gullibility heuristic hypothesis, we generated a competing explanation, one where the anchoring and adjustment heuristic was to be explained via the default heuristic that people should believe what they perceive when they are under load. But this explanation is not available to explain these self-generated anchoring and adjustment data. The endogenously created anchors aren't perceived at all. Consequently, I take it that the gullibility heuristic isn't a sustainable hypothesis for that heuristic's domain was exogenous stimuli, yet the same effects occur when using endogenous stimuli.

One might want to object to the above reasoning on the grounds that introspection is a kind of perception (one who subscribed to a Higher Order Perception theory of consciousness could do so in a non-ad-hoc fashion, Lycan 1987, 1996). Thus, one might think that the self-generated anchors are, in a sense, perceived, and thus the gullibility heuristic can proceed unscathed. But this would be quite an odd position to hold, since presumably we only have two ways of 'perceiving' stimuli inside our body: through proprioception and through introspection. This case does not clearly seem to be a case of proprioception. Proprioception is the internal perception of one's body, not mental states. Moreover, even if one takes introspection to be a form of perception, this would be a very odd type of introspection. Paradigmatic cases of introspection are cases where one is searching for an answer that one already knows that one knows, not producing one ex nihilo. So it strikes me as very odd to suppose that we'd turn the gullibility heuristic inward to

ourselves. Although I find this objection a bit specious (for it loses the impetus behind positing the heuristic in the first place and I don't like thinking about introspection as perceptual), there is still some further way to quell the proponent of the gullibility heuristic who like to see introspection as perceptual. In the next subsection, I will give my last argument against the gullibility heuristic objection

4.2.2: Evidence against the Gullibility Heuristic Part 2: Capgras Syndrome

Indirect evidence against the gullibility heuristic story can be gleaned from certain neuropsychological observations. A heuristic story would predict that neurological damage shouldn't cause a dissociation between acceptance and rejection, yet it seems that people with Capgras syndrome show such a dissociation. Unpacking this claim is the goal of this subsection.

The Spinozan model is an architectural processing model, one where there are two separable processes: the passive understanding/believing process and the active rejecting process. As such, the two processes should be dissociable. Architectural processes can be selectively inhibited by different types of neurological damage. Contrarily, heuristics (regardless of whether they are explicitly represented or not) are not the types of things that have been observed to be selectively inhibited by neurological damage. Thus, if rejecting a proposition is part of a different mental process than understanding/believing a proposition, then we should expect that the faculty of rejection can be impaired without impairing the 'faculty' of understanding/believing.[92]

I'd like to suggest that Capgras Syndrome gives us an example of people of a group of people who are able to form new beliefs, but not negate beliefs. Capgras syndrome

[92] As my scare quotes imply, I find talk of a faculty of rejection unobjectionable, but talk of a faculty of belief to be misleading.

patients have a condition that appears to selectively knock out one's ability to reject beliefs, thus providing evidence that the faculty of rejecting a proposition is separate from the faculty of accepting a belief.

Although the etiology of Capgras is disputed, a popular hypothesis is that it is caused by a lesion generally in the right posterior regions of the right hemisphere (between the fusiform gyrus and the amygdala, see Hirstein and Ramachandran 1997). The suffers of Capgras generally have delusions where they believe that their loved ones have been replaced by duplicates, as if they had been invaded by body snatchers. One popular hypothesis of the phenomena is that the sufferers have their emotional areas dissociated from their face perception areas. Thus, when a Capgras patient looks at (e.g.,) her spouse, the affect generally associated with such experiences is missing. Consequently, they conclude that their spouse has been replaced by a double.[93] However, the syndrome needn't be interpreted as merely a perceptual/affective problem.

A persuasive explanation of the syndrome is that Capgras "is the result of a secondary rationalization process on the part of the patient, to support their belief that the original person has been replaced" (Frazer and Roberts 1994). The idea that the syndrome is not essentially a perceptual problem, but instead is a problem of belief fixation is echoed in Berson: "Capgras' syndrome, then, is not a perceptual problem, an illusion, a hallucination, a misrecognition, or an autoscopic phenomenon. It is a problem of belief, a delusion" (1983, p. 971). With this interpretation in hand, we can readily see the dissociation that we need to support the Spinozan view and disarm the gullibility heuristic hypothesis. The Capgras

[93] Many Capgras patients also report that they believe that there are existing doubles of themselves (Berson 1983). Presumably they are gathering this belief from the quite quotidian experience of looking in the mirror, seeing their face and not receiving the normal affect that comes with such experiences.

sufferers have their inferential apparatus left somewhat intact, but are unable to incorporate any disconfirming information into their thought process. In essence, they are unable to reject what they think, regardless of any counterevidence present that should disconfirm their delusions. Once they see their friends and family as doubles no evidence presented to the contrary can persuade them otherwise.

This type of dissociation between acquiring beliefs and being able to doubt beliefs is predicted on the Spinozan view. The gullibility heuristic cannot predict such a dissociation, for heuristics are generally not disabled by specific neurological damage.[94] Thus, it appears that the neuropsychological data supports the Spinozan theory over the gullibility heuristic.

In sum, the arguments given should cast strong doubt over a heuristic solution to the problem of belief fixation. Since the gullibility heuristic is posited to only appear under load, it looks like a dubitable posit to begin with. Additionally, the heuristic solution alone does not seem to have the resources to explain the self-generated anchoring and adjustment effect; we need the Spinozan apparatus in order to explain it. Moreover, the Spinozan hypothesis, but not the heuristic hypothesis, can predict the types of deficits we see in Capgras Syndrome. Lastly, even if one were to ignore all of these arguments, the gullibility heuristic still can't explain the belief perseverance effects. With that I set aside the gullibility heuristic and shall turn my focus to other troubling objections.

4.3 The Informativeness Objection

[94] None of the preceding should be taken to imply that the ability to reject propositions is neurally localized in the sense that (e.g.) face perception may be. Nor is any this is meant to imply that Capgras sufferers have a clean dissociation, with just the ability to reject propositions knocked out. Since Capgras is generally caused by neurological damage, the damage tends to be diffuse and thus other abilities are also damaged (e.g., spatial memory, Christodolou 1977). Mother Nature may be a mad scientist, but she's certainly not a careful one.

Some theorists have proposed that people do have the ability to contemplate propositions without believing them, but only when the propositions are informative when they're false and not when they're uninformative when they're false (Hasson et al. 2005). If this were the case then most of the exogenous evidence used against the Cartesian view would be undercut.

Hasson et al. gave participants statements that were paired with faces. Participants were told that some of the statements were true and others false. Some of the statements were informative when true but not when false (e.g. 'this person walks barefoot to work'), some were informative when false but not when true (e.g., 'this person owns a television'), some were informative when either true or false (e.g., 'this person is a liberal'), and some were uninformative when both true and false (e.g., 'this person drinks tea for breakfast'). During the learning phase participants were instructed to memorize the statement/face pairs for later testing. Additionally, for some face/statement pairs participants were put under cognitive load.[95] In the testing phase participants revisited the faces and were asked to determine whether the accompanying statements were true or false of the person whose face they viewed.

For statements that were uninformative when false, the Spinozan prediction held: interruption had no effect on the recollection of true statements, but interruption increased participants' tendency to report false statements as true. On the contrary, interruption had no effect on statements that were informative when false. For these statements, interruption

[95] The load was induced by another tone task. The participants would hear a tone and they were instructed to detect whether the tone was high pitched or low pitched and then push a button corresponding to the pitch.

affected true and false statements equally. That is, for informative-when-false statements, participants remembered true and false statements equally well regardless of cognitive load.

This appears to be a decidedly anti-Spinozan datum, for it seems to show that people do have the ability to withhold assent from propositions when those propositions are informative when false. The experimenters write, "These results support the idea that the effect of resource depletion on the encoding of falsity ultimately depends on whether or not the proposition's false version is informative" (Hasson et al. ibid, p. 568). If their hypothesis is correct, then at best the Spinozan hypothesis's scope would be severely restricted. However, there is good reason to resist their conclusion.

First, it is important to note how odd the consequences of the informativeness hypothesis are. If the hypothesis was true, then people wouldn't be able to contemplate a proposition without believing that proposition *when the proposition is uninformative when false*. People could only contemplate without believing when they are thinking about propositions that are informative when false. This situation is theoretically untenable. Suppose that you encounter a proposition, P. If not-P is informative, then you will be able to contemplate P without believing it. However, in order for you to determine whether not-P is informative, you must first parse and consider P. But what happens when you consider P? When you're considering P, do you believe P or not? In other words, what relation do you bear to P before you have figured out whether not-P is informative? When first considering P you either believe it or you don't, but the informativeness hypothesis can't account for this fact because whether you believe it or not is based on whether the proposition is informative or not.

It seems overwhelmingly plausible that before a person can determine how informative a proposition is, the person must first entertain the proposition (though, of course, a person needn't consciously consider it). Consequently, it appears that the informativeness hypothesis must entail that people can withhold assent regardless of the (subjective) informational content of a proposition. This would in turn imply that after one has withheld assent, one goes and marks propositions as true *only when they are uninformative when false*. This is quite an odd situation indeed. The informativeness hypothesis dictates that people believe propositions after they've considered a proposition they've been told is false when that proposition is uninformative-when-false. How could such a situation come about? How would the mind possibly evolve such an odd processing system?

So, prima facie, we have good reason to be skeptical of the informativeness hypothesis. A much more coherent hypothesis would be the Cartesian hypothesis that we can contemplate without believing across the board. But all of the data we've considered (including Hasson et al.'s) speaks against the Cartesian hypothesis. What are we to do?

I suggest we hold that we take seriously the fact that the Spinozan hypothesis is true. One could reasonably hold the Spinozan hypothesis even in light of Hasson et al.'s results because their study was inherently flawed. Consider being a subject who has just seen two sentences, both of which you were told were false, one that is uninformative when false ('this person drinks tea for breakfast') and one informative when false ('this person owns a television'). Why would we be more apt to remember that the latter was false? Perhaps because the latter is more shocking and vivid. When we encounter abnormal situations (one might think: who doesn't own a television?), we are more apt to think longer and harder

about the abnormal situation.[96] Finding out that someone doesn't drink tea for breakfast doesn't really get one's mental juices flowing but finding out that someone doesn't own a television immediately raises some questions (e.g., is this person a Humanities professor? A Communist? Is she poor? Does she live in the woods?). Unsurprisingly, the more you think about something, the more you are apt to remember it. The subjects in this study were probably thinking about the statements that were informative when false for a much longer amount of time than they were the statements that were uninformative when false. Participants were probably considering the situation where one doesn't own a television for longer than they would consider the situation where one doesn't drink tea for breakfast (certainly the former would startle undergraduates more than the latter and of course undergraduates were the participants in the study). Accordingly, they would perseverate on the thought that contains the concepts DOES NOT OWN A TELEVISION, more than they would meditate on DOES NOT DRINK TEA FOR BREAKFAST, thus they would be more apt to remember the former than the latter. Seen in this light, Hasson et al.'s data tells us nothing about the processing of belief per se.

One last reason to think that my above explanation is correct: the informativeness criterion coincides with the ease of imagining a situation. When one considers someone who doesn't drink tea for breakfast what comes to mind? There is no concrete mental image that occurs. However, when one considers someone who doesn't own a television, then many mental images pop up (try this on yourself). In fact, one can see the difference in these

[96] There is some evidence that deals with this line of thought (see Brigard et al. 2009; Mandelbaum and Ripley ms.). The main thesis of the latter paper is that people have a belief that when a norm is broken, an agent must have broken the norm. The idea is that one gleans more (sometimes false) information about a person's mental states when they break norms than when they follow norms. If you see me on a tuxedo at a fancy wedding, then you don't learn nearly as much about my mental states as if you see me in a tuxedo at the beach.

statements as on par with the difference between the abstract and concrete innuendo effects. In studies of the perseverance of innuendos (e.g., Wegner 1984) we find that innuendos make a deeper impression when they are concrete rather than when they are abstract. People can more easily ignore innuendos that are abstract (e.g. 'Audrey is not sour') than they can for innuendos that are concrete ('Audrey did not rob Toys R Us'). Presumably this is because 'not sour' can be immediately translated into 'sweet.' Moreover, we know that people will flip negative statements into the equivalent positive statement whenever possible (see Wason and Johnson-Laird 1972). One can easily paraphrase and flip the abstract statements, but how could one do the same for the concrete statements? What comes to mind when I tell you that Audrey didn't rob Toys R Us? Was she at home sleeping? Did she attempt to rob it but was foiled by the Pinkertons? In sum, the concrete statements 'stick' because it's hard to envision a particular situation that holds when the statement is false. The difficulty of envisioning does not occur in the abstract statements because they have a quick negative counterpart.

Similarly, the informative statements in Hasson et al.'s studies can be easily envisioned when negated. When considering that this person doesn't own a television you may immediately think of a person living in a log cabin in the woods (or perhaps you envision someone reading, or a big old-timey radio, or a television that's been turned into a diorama, like mine). However, when I tell you that this person doesn't drink tea for breakfast what is the first thing that comes to mind? Do you envision a person sitting at a table with no drinks? Do you envision a coffee cup? The uninformative situations are hard to visualize when false. Thus, it is no wonder that people have a harder time remembering the veracity of uninformative statements than the veracity of informative statements. People will think about the latter more often and will thus be able to answer more correctly. Seen in this light,

Hasson et al.'s results tell us nothing about the relation between contemplation and belief per se, and do not cast doubt on the Spinozan hypothesis. In fact, in order to explain their results one needs the Spinozan hypothesis in order to explain why participants represent uninformative statements as true even when they are told they are false. As opposed to attacking the Spinozan hypothesis, the data collected in Hasson et al. helps to support Spinozan view.

4.4 Dretske's Objection from Non-Conceptual Content

Fred Dretske has posed the following objection to me: he argues that prior to central cognition, perceptual processes contain propositional information that is non-conceptual (see, for example, Dretske 1993, Fodor 2009). Since the information is non-conceptual it can't possibly be believed because one can't believe that X is an F without having the concept X. So, the objection goes, I either have to a) claim that you can believe what you can't conceptualize, b) deny that there's non-conceptual content, or c) claim that my view only applies to mental representations that are post-perceptual.

The first option seems the most unpalatable. Perhaps it only seems incoherent because I've been assuming a Language of Thought picture. Maybe a really nifty connectionist network could model a system that believes that P without it being able to conceptualize P. If one is willing to allow that we can represent P without having the concept P, then maybe we can believe that P without having the concept(s) P. There is some precedent for such a view in the literature. Sperber (1985) has held a similar thesis, arguing that people can believe statements they don't understand (but see Recanati 1997 for a subtle critique). Assuming you can't understand what you can't conceptualize (which seems right), maybe we could extend Recanati's arguments. But on second thought, I'd prefer to not be led

down this path, for I can only barely understand what it would mean to say that someone can believe that P without having the concept P.

The second option seems a bit better. Perhaps we can't represent what we can't conceptualize so perhaps there can't be non-conceptual content. Reasonable (ish) folks have held such a view (e.g., McDowell 1992) and I suppose I could embrace it. But, for reasons unrelated to the issues at play here, I'm decently impressed by the arguments in favor of non-conceptual content. At worst, I'd like to stay neutral on the non-conceptual content debate, at least for now. Since I'm not willing to argue against the existence of non-conceptual content, I'd better take the third horn and restrict the Spinozan view to just post-perceptual propositions. This isn't necessarily a drawback. One can consider this a discovery as opposed to a problem. We could interpret Dretske's objection as delineating the explanatory reach of a Spinozan theory, which just helps to specify the Spinozan hypothesis. In fact, taking this route would seem to have some very nice consequences. It would dictate that mental states that are the least intuitively thought to be beliefs, information bearing non-conceptual states, aren't beliefs. Perhaps Dretske's suggestion should just be welcomed.[97]

4.5 The 'Why Aren't These States You Call Beliefs Just Aliefs?' Objection

Short answer: because there are no aliefs. If you are convinced of this, feel free to skip this section. If not, let me introduce you to the notion of alief before arguing that we should discard it.

[97] One might think that a more difficult question for the Spinozan to face is what to say about intramodular propositions (assuming there are modules and that some contain propositions). I suspect this problem is actually a red herring, for the canonical intramodular propositions that have been offered (e.g., the proposition that objects don't lose their size as they recede from the observer, Fodor 1983) look to be belief-like. The problems with accepting these propositions as beliefs has little to do with the Spinozan view and more to do with their lack of full-fledged inferential promiscuity. However, since I agree with Egan (ibid.) that most belief systems are fragmented, we shouldn't suppose that just because a proposition is not fully inferentially promiscuous, it's not a belief. For all we know (or better, for all I know) it may be the case that intramodular propositions are inferentially promiscuous inside their home module.

In a series of recent papers Tamar Gendler has put forward the provocative idea that there exist a set of mental states, aliefs, which play an integral role in our mental economy yet have hitherto been undiscovered (Gendler 2008, 2009). Gendler's work on alief is intriguing and far-reaching; if her hypothesis proves true, it would call for a reconceptualization of many well-known, though perhaps not well-understood, psychological phenomena. Consequently, her proposal deserves serious attention and scrutiny.

I do not think that the notion of aliefs survive such scrutiny. The structure of my argument proceeds as follows: first I will go through Gendler's characterization of aliefs. I will argue that the properties of aliefs only separate aliefs from beliefs if aliefs are understood as essentially associative. In the second section, I will use one of Gendler's examples to argue that aliefs (at least sometimes) must be propositional and are thus not purely associative. I will then end the paper with some positive suggestions of what the discussion of the non-existence of aliefs can teach us about beliefs.

4.5.1 Characterizing Aliefs

Before we can adequately discuss the ontological status of aliefs, we must first get clear on what exactly aliefs are supposed to be. This task is a bit difficult, since the notion of alief is quite slippery. Consequently, I will rely heavily on Gendler's actual words in order to explicate the idea. Gendler writes that an alief is a mental state with

> associatively linked content that is representational, affective and behavioral, and that is activated—consciously or nonconsciously—by features of the subject's internal or ambient environment. Aliefs may either be occurent or dispositional. (Gendler 2008, p 9).

Let's unpack this quote, for it provides a somewhat problematic characterization. This is because the given characterization doesn't separate aliefs from other psychological states we already countenance. Each one of the properties mentioned in the characterization can be

used to describe beliefs (or, with a little work, concepts). Gendler starts her characterization of aliefs with the phrase 'associatively linked content,' which makes it appear as if she intends aliefs to only have non-propositional content. However, when Gendler explains what she means by associatively linked content, the meaning becomes less clear. On the question of association Gendler states that aliefs are made of "a cluster of contents that tend to be co-activated. The contrast here is with discrete contents that fail to be linked through such an association (Gendler 2008, p. 9)." I find this gloss a bit puzzling. It seems as if she is saying that aliefs aren't *necessarily* associative (hence the use of 'tend'). However, if that's what she means, then aliefs are just like beliefs and concepts in general. The concept of SALT tends to be co-activated with PEPPER, as does CAT and DOG; likewise the belief THAT IS A TIGER IN MY BED tends to co-activate other behavioral and cognitive states (like a fear response and a belief that one should run away). Thus, prima facie, in order for aliefs to be interestingly different than beliefs (or concepts), aliefs must be essentially associative (a point we will return to below). This is because the rest of the properties enumerated in her characterization of aliefs untendentiously apply to beliefs, a point I'll return to immediately.

For example, another property of aliefs is that they may occur unconsciously. Thus, a person may hold the particular state without knowing that one is in that state. Yet it is untendentious that most mental states can be tokened unconsciously (for example, in the prototypical priming paradigm one tokens a concept unconsciously.) Additionally, beliefs and other propositional attitudes can also be tokened unconsciously (see, for example, the role of desire in Freudian psychology, the role of belief in cognitive dissonance explanations

or in explanations of implicit racism).[98] Thus the property of being unconscious does not separate 'aliefs' from beliefs and other mental states.

Gendler also characterize aliefs as having contents that are representational, affective and behavioral, but these additional properties don't help distinguish aliefs from good old beliefs (and concepts). For example, imagine the output of a Fodorean visual module (or better, of the last module in the visual system whose output goes to central cog, see Fodor 1983). Let's say the output is LO, A PANTHER. This output a) is representational, b) has an affective component (presumably thinking of ambient panthers normally causes sweating, fear etc.) and c) is associated with the readying of the fight or flight routine, which is a behavioral component. Yet, LO, A PANTHER is just another run-of-the-mill belief.

It's not just beliefs that have representational, affective, and behavioral components. Take the concept PENGUIN. This concept appears to be closely associated with the following information: PENGUIN is pronounced (pĕng'gwĭn). Don't you think that when we think of penguins we are quicker to say 'penguin', spot penguins, mistake things for penguins? Merely tokening the concept PENGUIN readies us for penguin-related behaviors, which is bad news for the alief aficionado. Gendler writes that "aliefs don't involve the execution of these motor routines; it merely involves their activation (Gendler 2008, p. 11)." However, the same holds for all types of quotidian mental states. Tokening PENGUIN (or believing, THERE'S A PENGUIN) doesn't involve executing motor commands either, but it does ready them; if it didn't why would we (e.g.,) be faster at lexical decision tasks involving 'penguin' after we've tokened PENGUIN (n.b., the tokening needn't be conscious of

[98] For example, one way to explain the workings of implicit racism is to posit that the implicit racist harbors an unconscious belief that (e.g.,) Caucasians are superior to African Americans.

course)? Lastly, tokening PENGUIN is generally associated with some affect too (doesn't thinking about penguins make you feel all warm and fuzzy?). Once again, the hallmark properties of alief appear to be identical to the properties of other canonical mental states.

Next Gendler discusses the activation conditions of an alief: "It may be activated by features of the subject's internal or ambient environment" (Gendler 2008, p 9). Needless to say, but the same holds for beliefs (and concepts). One can token the belief I AM HUNGRY by sensing one's bodily states or by certain cues from the ambient environment (e.g., hearing one's stomach rumble or having someone point out that you happen to be mechanically eating potato chips which you are known to dislike). To round out the properties, Gendler mentions that aliefs may be either occurrent or dispositional. Of course, it's not particularly tendentious to think that the same is true for beliefs.

It should be clear from this short discussion that aliefs seem to have the same properties as other mental states. As such, we appear to have no need for a new category of mental states, for the explanatory burden that aliefs are supposed to relieve can be carried out by psychological entities we already countenance. If aliefs are to do any work for us, then aliefs must somehow differ from beliefs. As just canvassed, in order for aliefs to be sufficiently distinct from beliefs, aliefs must be *essentially* associative. In the next section I will analyze one of Gendler's alief examples and show that aliefs must be propositional. Once we see that aliefs must be propositional, we can safely infer that the states Gendler terms 'aliefs' are just beliefs and hence we discard with the notion of aliefs.

4.5.2 Why Aliefs Cannot be Essentially Associative

I have two arguments for why aliefs cannot be essentially associative: one argument based on what I'll term 'binding'[99] and the other based on inferential promiscuity. To see how these arguments work, let's consider Gendler's alief-based explanation of Paul Rozin's poison experiment (Rozin et al. 1990, put to use in Gendler 2008). In Rozin's experiment participants are shown two empty bottles that are subsequently filled with sugar. The experimenter then shows the participant two labels, one saying 'Sugar', the other saying 'Sodium Cyanide.' After reading the labels, participants are more hesitant to drink from the bottle with the 'Sodium Cyanide' label. Gendler concludes that though the participants believe that both bottles contain sugar, they alieve that one of the bottles contains sodium cyanide.

The problem is that the putative alief looks to be propositional and so the alief explanation of the situation won't due. To begin let's look a bit closer at the Rozin example Gendler uses. Gendler claims that the content of the alief at work is "CYANIDE, DANGEROUS, AVOID" (Gendler 2008, p.15). However, we should pause to ask ourselves: what is this alief specifying should be avoided? To put the point another way, when I token the alief with content 'CYANIDE, DANGEROUS, AVOID', what am I thinking? If I'm just tokening these concepts in succession (which is what her 'associative state' talk implies), then why would I show any behavior whatsoever towards the bottle and not, say, the window, my left foot, or the experimenter's forehead? Since the behavior is bottle specific, the putative alief must somehow *bind* to the bottle, or else participants wouldn't show the

[99] A warning to the reader: my use of binding is not the use of binding at play in say cognitive neuroscience discussions of the 'binding problem' as it's used in, e.g., visual perception. Although these are different senses of binding I find the use of the term to be helpful at getting at the underlying idea. However, there is precedent for my usage: Roskies (1999) sees the problem I'll describe and the traditional binding problem as structurally similar. I apologize for any confusion this might cause.

avoidance behavior toward the bottle. Thus, the alief must have a content more akin to THAT [demonstrative standing in for the bottle] DANGEROUS CYANIDE AVOID. But are the participants just thinking the concepts THAT DANGEROUS CYANIDE AVOID, one after another, with no syntax as it were? If so, then why would we avoid that particular bottle? Instead, it seems like the participants must be thinking something like THAT IS DANGEROUS CYANIDE, AVOID IT.[100] To put it differently, if these aren't propositional thoughts (e.g., if we just think DANGEROUS CYANIDE AVOID without any propositional structure), then how can we make inferences from the alief DANGEROUS CYANIDE AVOID? How does it happen that we infer e.g., that it's still the same dangerous poison even when it's not in my ambient environment, even if I put the bottle behind a curtain, or close my eyes?[101] The present problem is that the content of the alief must somehow bind to the bottle and the associative content that Gendler specifies for the alief has no way of attaching to the bottle as opposed to anything else. In order to bind in the right way, the content needs to be structured. And, as I'm told Kant pointed out to Hume, associative content can't provide the right type of structure (Fodor 2003). This is what I was terming 'the binding argument.' The contents will only eventuate in predictable behaviors (e.g., bottle-avoidance behaviors) if the contents are bound in the right way. Propositional structure would allow for such a binding, but purely associative structure won't do.[102]

---

[100] Of course, they could just think THAT IS CYANIDE with CYANIDE being linked to DANGEROUS, which itself would be linked to avoidance behaviors.

[101] FINST style explanations (Pylyshyn 1989b) are of no use here for the FINST trackers break down after prolonged occlusion and presumably one would still avoid the bottle after somewhat prolonged occlusion.

[102] N.b.: you can't fix this problem by adding a fourth element to the content, such as an iconic representation of the bottle. Say the alief had the content CYANIDE, DANGEROUS, AVOID, PICTURE (where picture stands in for an iconic representation of the bottle). In such a situation it would still be a mystery why anyone would avoid the bottle because these would be four separate thoughts, albeit thoughts that sequentially followed one another. If you're having trouble seeing the difference perhaps the following example will prove illuminating. Imagine we have two cognizers, one who tokens the thought SEXY WILDEBEAST (a single thought with an

Moreover, pure associative chains don't allow for inferences, but the putative aliefs do appear to allow for inferences so they must be propositional, in which case these states are truth evaluative and appear to work just like beliefs. In other words, the aliefs seem to be inferentially promiscuous, but if aliefs are essentially associative then they should be inferentially dormant—after all, one can't make (truth-preserving) inferences from associative chains.

To see how the putative aliefs can be inferentially promiscuous, imagine that right after you take part in the Rozin's study, you are asked a follow-up question about whether other folks would drink from the bottle with the cyanide label. In this case you'd probably infer that others won't want to drink from the bottle. (Perhaps you go through an unconscious chain of reasoning like *that bottle contains poison, people don't like drinking poison, so my people won't like drinking from that bottle*). In short, we should expect people to infer from THAT IS CYANIDE, DANGEROUS, SO AVOID IT, to other semantically related (and under the circumstances, reasonable-ish) thoughts, such as thinking that others will want to avoid the bottle labeled cyanide, that the bottle would still be labeled 'cyanide' even if the room was a different color, that the bottle will keep its contents even if it's lifted off the ground, etc. There are an infinite amount of quotidian inferences we'd expect the participants to make but these inferences can only be made from propositional states. Hence there must

adjective noun structure) and the other who tokens SEXY followed by a tokening of WILDEBEAST (two separate thoughts). These are two very different cognizers; the first one clearly has some odd sexual proclivities, whereas the second one just appears to be someone lost in a stream of consciousness. We can predict a decent amount of the first person's behavior from knowing that the person tokened SEXY WILDEBEAST (for example, you probably wouldn't want to let him babysit your wildebeest), but we can't predict much of anything at all about the second cognizer. If aliefs were essentially associative, then alief contents would parallel our second cognizer. But this can't be right, because we *can* predict the behavior of the participants in Rozin's experiments: we know they are apt to avoid the poison.

be belief-like propositional states in play. But if we already need beliefs in play in order to explain the phenomena, then why bother positing any aliefs at all?

My hunch is that Gendler wants to posit aliefs in order to save beliefs for epistemologists. I think Gendler is committed to two other ideas that motivate her picture of aliefs: the idea that people don't have contradictory beliefs and the idea that beliefs are consciously introspectable. The latter has already been dealt with extensively in section 4.1 so although I'll discuss it a bit more, I'll keep the discussion short. Before we get there, let's take a moment to discuss the status of contradictory beliefs.

The idea that people aren't overrun with contradictory beliefs is one that plays an important role in epistemology (e.g., in specifying a Coherentist descriptive account of justification or in Bayesian models of belief updating). Consequently, Gendler does not want to say that the person believes e.g., that the bottle does and does not contain cyanide because she doesn't want to attribute contradictory beliefs to people.[103] However, it's clear that people often just do hold contradictory beliefs, and these beliefs needn't all be unconscious. Go to your local soup kitchen and I bet you you'll find this phenomenon quite quickly. A not wholly uncommon phenomenon is for people with low-self esteem to partake in some altruistic volunteering. Frequently enough one will think 'I'm a terrible person' and this will cause them to, e.g., volunteer at a soup kitchen. During this volunteering the person will think 'I'm a wonderful person' while simultaneously believing that she is a terrible person. Along similar lines, Milgram elegantly noted that most people seem to believe that murder is wrong and that murder isn't wrong, often at the same time (Milgram 1974, p. 7).[104]

[103] Although she never says this in her papers, she did admit as much in the question and answer session after her talk at the 2009 SPP.

It's not just 'the folk' that hold contradictory beliefs; academics also do so all the time. Imagine a philosopher who has just completed the oral defense of her dissertation (ahem). It is quite normal for one to think that one is an idiot right after the oral defense because one couldn't adequately respond to all the questions posed during the defense. Now imagine that a short time later the same philosopher is sitting on someone else's committee during his defense. It is quite natural for the philosopher to now think that she is quite smart because the poor graduate student in front of her cannot answer her questions.[105] However, that doesn't mean that the first belief went away. The best way to describe this situation is as one where the philosopher believes both that she is smart and that she is an idiot. We can then explain which belief is active by appealing seeing which belief would be more salient based on the impinging external stimuli (along with whatever internal states happen to be active).

Lastly, note that a very natural way of describing people's responses to vague cases is to appeal to people's contradictory beliefs. One doesn't have to be a dialethist to suppose that most people, when queried, will end up believing that (e.g.,) 12:15 is both noonish and not noonish (in fact, there is solid experimental evidence that people not only hold simultaneous contradictory beliefs in vague cases, but also that they are willing to assert the contradictory beliefs in vague cases; see Ripley 2009). Thus, we should not posit aliefs just to save us from

[104] "The force exerted by the moral sense of the individual is less effective than social myth would have us belief. Though such prescriptions as "Though shalt not kill" occupy a preeminent place in the moral order, they do not occupy a correspondingly intractable position in human psychic structure. A few changes in newspaper headlines, a call from the draft board, orders from a man with epaulets, and men are led to kill with little difficulty" (Milgram ibid. p.7).

[105] This may be the most awkward paragraph I have ever written. Which is saying a lot (I was an editor of an adult literary magazine in college and that was less awkward than this—strike that, writing this parenthetical takes the cake). While we're here though, it may be for pointing out that the example in the text was the inspiration for Nisbett's research on the fundamental attribution error.

ascribing contradictory beliefs to people because we will have to ascribe contradictory beliefs regardless of how the alief debate works out.[106]

The relation between beliefs and introspection is a bit thornier. If I'm right and these aliefs are just beliefs, then why don't people report having these beliefs? Why is it that some of our beliefs are consciously available for report (e.g., the belief that I'm now typing) and some aren't? This is a difficult question with a fine pedigree and my answer to it was given above in section 4.1. Instead of rehashing those issues, I now want to take a moment to peruse other areas of psychology that are predicated on one's lack of introspective access to one's beliefs, for such a discussion will illuminate our discussion of aliefs.

Let's start with an example from the 'insufficient justification' paradigm of cognitive dissonance theory (see, e.g., Festinger 1957). Suppose a guy arrives at college and doesn't know exactly how he feels about fraternities so he tentatively rushes one. Because a body in motion tends to stay in motion, he halfheartedly continues to pledge, putting in little effort. Throughout the semester the pledging becomes more and more effortful. What we find in these situation is that the more effortful the pledging, the more people report liking the fraternity. This is because the fraternity brother has to justify the effort he has put it: he reasons that the more effort he puts in, the more he must like the fraternity. He justifies his effort by changing his opinion. This effort justification works in a predictable and reliable fashion (hence the experimental paradigm named 'the effort justification' paradigm). One

[106] Unsurprisingly, my preferred analysis of the Rozin experiment is that his participants hold contradictory beliefs: they both believe that the bottle contains cyanide and they believe that the bottle doesn't contain cyanide. The reason they believe the bottle contains cyanide (even though they have good evidence to the contrary) is that people have tokened a thought of the form THAT BOTTLE HAS CYANIDE when they saw the cyanide label on the bottle and thus acquired the corresponding belief.

infers from the premise 'I put a lot of effort in' and the premise 'I'm not a schmuck, and only schmucks put a lot of effort into something they don't like' to 'I must really like this.'

The points to keep your eye on are that a) the whole chain of inferences proceeds unconsciously, b) the premises themselves sure look like beliefs and c) the premises were acquired unconsciously.[107] Also worth noting: this type of reasoning is totally ubiquitous and is the backbone of a very robust, successful, and established psychological research project: cognitive dissonance theory.[108] Moreover, this sort of reasoning is part of a healthy psychological immune system and gets stunted if any of it is brought to consciousness. In essence, the reasoning at play in the explanations of the insufficient justification paradigm refers to states that are quite clearly beliefs, yet seem to have all the same properties as aliefs. These states are full-fledged beliefs because they are informationally promiscuous (after all they are serving as the premises in inferences!), though they are not available to conscious report. So, even though Gendler may want to posit aliefs in order to save beliefs from being unintrospectable and not particularly rational, we need to countenance such belief states

[107] Maybe his parents told him he's not a schmuck. Assuming so won't change the point in the text.

[108] This is just one of a giant group of examples that work the same way. I haven't the space (or ability) to even try and list (or even characterize) the group. Instead I'll just mention one more fun example, this time stemming from attribution theory as opposed to classical dissonance theory, just in case one has a particular hatred of dissonance theory. Storms and Nisbett (1970) ran an experiment on insomniacs, the partial goal of which was to help their insomnia. The insomniacs were split into two groups. Both groups received a pill before they went to sleep, one group was told that the pill would cause arousal and the other told that the pill would reduce arousal (both pills were placebos). Their fascinating result was that the group that was given the pill that would exacerbate arousal ended up getting to sleep quicker and easier than the group that was given the pill that was supposed to reduce arousal. The reason for these results is as follows: both groups went through unconscious chains of inferences. The first group felt aroused before they went to sleep (which is why they're insomniacs in the first place) but then attributed this arousal to the pill and not to the sleeping situation, which alleviated their sleep-based anxiety and allowed them to go to sleep quicker and easier. The latter group (the group that had been told that the pill they were taking would assuage arousal) had their insomnia exacerbated because they too felt aroused before sleep but since they took a pill which was supposed to alleviate arousal they then 'over-attributed' their arousal to the sleep situation (unconsciously reasoning that since they should be feeling less arousal than normal and yet are still feeling aroused, sleeping must really be an anxiety inducing endeavor). Note that these fecund explorations only get off the ground by granting people a host of unconscious beliefs and chains of reasoning.

anyway. Hence, not only do we have no need for aliefs, but aliefs couldn't save us from the irrational aspects of belief anyway.

4.5.3 The End of Aliefs

I've argued that there are no aliefs. My argument was that either aliefs are essentially associative or else aliefs behave exactly as beliefs. However, aliefs cannot do the work Gendler asks of them unless they propositional, so aliefs are not essentially associative. But if they are not essentially associative, then they are beliefs. So there are no aliefs.

4.6 Conditionals, Liars, and Other Assorted Conundrums

In this section I'll cover a grab-bag of objections that have been put toward the Spinozan theory. The entities in the section don't really hang together in any natural way, except in so far that my responses to them will be brief.

4.6.1. Conditionals

Take your favorite conditional. Say it's 'if I bathe in the East River, then I'll be super clean.' My critic might say that she can wonder whether this conditional is true and she can also separately wonder whether its antecedent is true. The Spinozan theory says that if you entertain it, then you believe it. But what is it exactly that one believes when one tokens a conditional?

My short answer: I'm not sure. But I have some (mostly un-illuminating) things to say. First, when one feels like one is wondering about the truth of a certain proposition, whether conditional or not, what one is doing is forming a judgment as to whether the proposition is true or false. However, in terms of belief, the belief is already formed as soon as the proposition is tokened. As a consequence, many people will judge that they don't believe certain propositions that they do believe. All of which is just a long way of saying

that your introspective capacities can't discern what beliefs you've formed and what propositions you've judged as true. As we covered in 4.1, what you judge as true occurs at the person level, but what you believe is sub-personal. So, you can have the feeling of contemplating a conditional (just like you can with a declarative) and yet still not be withholding assent.

As to what happens when you token the conditional, my answer is unsurprising: you believe it. Of course, this is as vacuous as it is unsurprising. And it's vacuous for good reason: I have no idea how conditionals are processed in general, nor do I have much to say about what truth tables people use when in assessing conditionals. I wouldn't be shocked if people process different types of conditionals differently, applying different truth conditions depending on context. But none of that affects my answer here. You tell me how we process conditionals and I'll immediately turn around and tell you what exactly it is we believe when we believe a conditional.

4.6.2 Liars

Fact: we can contemplate the liar sentence. Many people spend their careers trying to figure out under what conditions the sentence is true. My critic may complain that the Spinozan says that everything you contemplate you believe, so then the Spinozan must say we believe the liar sentence. But how could anyone believe the liar sentence? We aren't even sure about what its truth conditions are.

I admit this is a problem. I don't exactly know what it means to say that one believes the liar sentence. But this doesn't bother me all that much because I'm not sure that anyone knows exactly what to say about the liar sentence. If forced to say something, I don't see why a dialethic treatment (Priest, 2006; Ripley ms.) would be any worse for the Spinozan than for

anyone else. The Spinozan already conjectures that people hold tons of contradictory beliefs, so taking dialethism on board wouldn't appreciably change her commitments.

4.6.3 Incoherent Sentences

It seems like people can entertain thoughts that don't really mean anything. Take the famous sentence 'Colorless green ideas sleep furiously.' What happens when one entertains that thought? Does one believe it? If so what are they believing?

Keeping with the theme of this subsection, let me reiterate my ignorance: I don't really know what to say about this case. My first instinct is to say that there is no proposition that corresponds to the thought COLORLESS GREEN IDEAS SLEEP FURIOUSLY. This sounds alright to me and perhaps this strategy can be expanded to the heterogeneous set of incoherent thoughts. Perhaps the thought THERE IS A CIRCULAR SQUARE has no corresponding proposition. But this does strike my ears as odd. I think I know the truth conditions for something being a circular square: it has to be a thing with no sides and four sides. Just because nothing has those properties does not mean that the sentence can't have truth conditions.

Maybe a better strategy can be had by conjecturing that people can believe propositions that they don't understand. This sounds a bit better to me. We do have some idea of what this scenario might look like: people often believe quite fervently in, at best, conceptually confused (and at worst conceptually incoherent) propositions. Additionally, there are respectable views of concepts that allow for this possibility. A conceptual atomist (e.g., Fodor 1998) holds that to have a concept is to be locked onto a property (the referent of the concept). Atomism thus allows for the very live possibility that people will have concepts with no beliefs whatsoever about the properties of that concept. For example, the conceptual

atomist could allow that I have the concept TREE, and thus can think about trees as such, without knowing anything about trees (e.g., without being able to identify them in any way). Now imagine a person who has just one belief about trees: that they are smaller than the galaxy. Do we want to say that this person understands this proposition? It seems a bit odd to say that. If you share this intuition, then maybe you'd be game to allow that people can believe propositions that you don't properly understand. Regardless, this objection strikes me as fairly exotic and I propose to drop discussion of it immediately.

4.6.4. Sub-propositional Mental Items

Critic: "There are many sub-propositional items that I can token, and these don't seem like they can be the appropriate objects of belief. For example, I can token CAT, but I can't just believe CAT."

Response: You are totally right, which is why the Spinozan theory only applies to propositional thoughts. Sub-propositional thoughts can't be believed, so, um, they can't be believed. I don't even know what it would mean to say someone believes CAT tout court. Whether one can entertain sub-propositional thoughts is an interesting question, one that, at this moment, I don't have much to offer. I suspect that one can, that people can just perseverate over one thought. Again, there are theories of concepts that certainly allow for this as a logical possibility. For example, conceptual atomism allows for the possibility of minds that consisted of only one concept, so that one could just cognize CAT all day and nothing more. Whether this is a nomological possibility or not is not an issue on which I have an opinion. Regardless, what I do know is that the Spinozan can side-step that question by just restricting the theory's scope so that it only ranges over full propositional thoughts. I hereby reiterate my commitment to so doing.

4.6.5 Logically Incompatible Propositional Attitudes

Critic: "Say I hope that I will wake up and be in Switzerland. It seems that the part of the reason I might hope such a turn of events would come up is because I don't believe it will. But the Spinozan theory says if I token a proposition then I believe it, which means that every time I hope that X will happen, I believe that X will happen. But surely this is absurd, for I wouldn't have hoped for X if I thought X was coming. So, there."

There indeed. This appears to be an awkward situation for the Spinozan, one that will need a little time to spell out. First let me get clear on what the Spinozan commitments by going through a side tour of some things the Spinozan has to say about Frege and then after that I'll return to the original complaint.

4.6.5.1 Frege's Thesis and the Fundamental Propositional Attitude[109]

Frege insisted upon distinguishing two dimensions along which mental events or states can differ from each other. The first dimension was with respect to their *force*, and the second dimension was with respect to their *content*. Two mental events or states differ along the first dimension if one of them is, say, a belief while the other is, say, a desire. Two mental events or states differ along the second dimension if one of them is, say, about the presence of a squirrel in the fireplace while the other is, say, about the heights of skyscrapers in New York. Frege's idea was that variation along one of these dimensions was independent of variation along the other. Two mental states could occupy the same location along the first dimension while occupying distinct locations along the second dimension, or vice-versa. For instance, you and I could both have a belief, but your belief is that the squirrel is in the

[109] I thank Ram Neta for the considerations that constitute this subsection.

fireplace, while my belief is that the Empire State building is taller than the Chrysler building. Similarly, you and I could both have some mental state with the content that the squirrel is in the fireplace, but your mental state is one of belief, whereas mine is one of desire. In this respect, Frege thought, variation along the dimension of force is independent of variation along the dimension of content. Furthermore, even if having some one of these mental states requires one to have lots of other mental states, it does not require one to have some other *specific* mental state. Frege's thesis can be summed up as follows: one can have a propositional attitude, A, with content C, without having any other specific propositional attitude.

The Spinozan theory of mind entails that Frege's thesis is false. A consequence of the Spinozan theory is that anytime I have a non-belief based propositional attitude with content P, I also must have a corresponding a belief with the content P. So, if I hope that P I also believe that P. In order for one to have the hope, desire, etc. that P, one must also have the belief that P. To repeat, in order for one to have any non-belief based propositional attitude one must also have a corresponding belief based propositional attitude with the same content. Thus, the Spinozan theory entails that Frege's thesis is false.

Frege's thesis has been very influential. Contemporary functionalists follow Frege by give no specific priority of the attitudes; that is, no attitude is deemed more basic than any other attitude. For functionalists, the attitudes all hang together in some yet unspecified way. Contrarily, the Spinozan view holds that belief is a more psychologically basic attitude than desire.

One can thus see belief as, in a sense, the fundamental propositional attitude. There have been some recent rumblings in the philosophical literature that point toward this

phenomenon. For example, in Gendler's discussion of aliefs, she points out that aliefs are ontogenetically and phylogenetically prior to other propositional attitudes (Gendler 2009, forthcoming).[110] She sees aliefs as fundamental mental relations. The only idea I wish to add to this insight that the distinction between aliefs and beliefs is a specious one (see section 4.5). It is belief that is phylogenetically and ontogenetically prior to our other propositional attitudes.

But of course, that not all I have to add because one might want to press the question that began this section. It's all fine and good to claim that Frege's thesis is false, but how do we make sense of someone hoping that p will come about if she believes that p will come about?

Let's use a test case. Say I really hope to get a job in South Dakota. In which case I might find myself thinking I WISH I COULD GET A JOB IN SOUTH DAKOTA. Well, if that's what I token, then all the Spinozan says I believe is I WISH I COULD GET A JOB IN SOUTH DAKOTA, in which case we get no contradictory state of affairs for all one would believe is that they wish that they could get a job in South Dakota. Note that tokening this

[110] One might wonder about the 'phylogenetic' claim there; in particular, one might wonder how it is that Spinozan beliefs evolved. I have wondered the same thing. Though the evolutionary evidence is mostly non-existent, I suppose there's no harm in engaging in some 'just-so' speculation (isn't that what footnotes are for?). Here's a 'just-so' story. It's not implausible to suppose that perception came on the scene before cognition. Some, like Prinz (2002), have even thought that cognition is an extension of perception, with concepts identified as percepts held in long-term memory networks. Now imagine that our perceptual faculties were by and large veridical. In such a case we can imagine that we'd save a lot of cognitive space by just taking the propositions given to us by perception as true. Thus, it might make sense for us to develop a cognitive system that just automatically believed the thoughts that passed through its mind. Automatically believing what you see is a reliable evolutionary strategy as long as one's perceptual systems are veridical and working correctly. Since we're assuming that our perceptual faculties were delivering veridical outputs, it is sensible that cognition would evolve in such a way as to take advantage of these preexisting structures.
If our perceptual faculties were delivering veridical percepts, the need to question these percepts would be obviated in most circumstances. In such a framework it would be preferable to not question the upshots of perception except in special circumstances. At some later point we then develop the ability to reflect on our own thoughts and reject some of them, but such rejection is an effortful endeavor. The rejection process is made effortful so that we don't waste all of our time rejecting the onslaught of (mostly true) propositions delivered to us from perception. My point here is not that this is the right story, but that some such story isn't unimaginable. And may I add: just so.

thought does not get you I BELIEVE I WILL GET A JOB IN SOUTH DAKOTA; to do that one would have to entertain the possibility that one will get a job there.

But my critic may not be impressed with this line of thought. Critic: "Yes, but say I wish for world peace, in which case I token THERE EXISTS WORLD PEACE and then hold the wish relation towards it. Well, here I've tokened the thought THERE EXISTS WORLD PEACE, so your theory entails I believe it. Which, may I add, is crazy." My critic has a point: for any propositional attitude I token, I just token a proposition and instantiate some relation toward that proposition. Wishing that p doesn't involve tokening one's WISH concept anymore than believing that p entails tokening one's BELIEF concept.

I think we can work around this problem. The way out of this puzzle is to note that we are right at the edge of where folk psychology and scientific psychology butt up against each other. I would like to think about this essay as one that attempts to take the folk term 'belief' and helps it along its way to psychological respectability. The states I've been calling 'beliefs' have the hallmark properties of beliefs: they are semantically evaluable (or if you prefer, they have satisfaction conditions), they are inferentially promiscuous, they can serve as (unconscious) premises in arguments to generate more beliefs (and knowledge), they display opacity, they are relational, and they interact with other mental states (most notably desire) in order to produce intentional action. Yet for all that they do lose what we might have pre-theoretically taken to be properties of belief. Pre-theoretically we might have thought that beliefs are introspectable. The beliefs we end up with are not introspectable and are closely tied to the foundations of thought (in the sense that if you entertain it, you bought it). As such they importantly differ from the folk sense of belief, the sense that, say, theorists use when discussing intentional action and practical reason.

I have no truck with the folk sense. I don't think it should be eliminated nor do I think that people should stop theorizing with it for certain purposes. But it's important for us to be clear what terms we are using where. I think the objection which started this section is actually mixing and matching folk terms with properly sanitized scientific terms. It may be a contradiction to wish that x whilst one believes that x will happen (though I doubt it), but there needn't be any contradiction between believing in the sense at use in this essay.

One may retort that although there might not end up being a properly sanitized notion of wish, there will probably be for desire. In which case the same problem might arise: how could I desire X when I believe X will happen? But I don't see any problem with this state of affairs. For one thing, I'd like to know a bit more of what the properties of properly sanitized desire are; for another, it seems perfectly fine to me to desire X and still believe X will come about. Right now, I seriously desire to have a cup of coffee and I believe I will—I don't see any logical incompatibility there and I hope neither do you.

4.6.6. Conclusion

One may not be convinced that the Spinozan theory is true, and my goal in this chapter has not been to argue that it is. Instead, I hope to have convinced you that the difficulties the theory faces are not insurmountable. If so, then perhaps the theory really has a future as a fecund research program. Along the way to arguing that it's a fertile program I've argued for a few other conclusions, most notably that we can't introspect our beliefs, that there are no aliefs, that I have offered a scientifically respectable notion of belief that still bears tight resemblance to the folk belief, and that Frege's thesis is false. I think it is probably best to wrap this up before my hubris eats me alive. To end our tour, in the final chapter I'll mention so of the more exotic points we've passed over. Then I'll take stock with

where we are, we're I'm planning to go next, and note some further observations and consequences of the Spinozan theory.

Chapter 5: Last Rites and First Approximations

After investing all this time into reading this, it's fair to ask: where have we ended up? In this final chapter I'll spend a little time gesturing in a few directions that I find interesting. This is in no way intended to be a comprehensive overview of all the commitments, consequences, and applications of the Spinozan view. Instead, I'll just guide you around the local flora and fauna that struck my fancy. I'll say a little bit about rationality, propaganda, the historical roots of the Spinozan view, and I'll speculate a bit about the picture of mind that I see emerging in contemporary cognitive science.

5.1 Rationality

As mentioned at feels like eons ago, at the outset of this essay (section 2.2), the Spinozan theory has some peculiar consequences for our conception of rationality. There I quickly discussed how central the Cartesian theory is to our views of what it is to be a person. The Cartesian view is intuitive in part because it's an integral piece of our picture of what it is to be a person. Pre- (and often post-) theoretically we see ourselves as creatures that can be conservative, weighing evidence judiciously and deciding what to believe. Upon reflection we might decide that we are not fully in control of what to believe because we are forced to believe what we have evidence to believe. When we abrogate control over our beliefs, we generally do so because we think we are responsive to a higher power: reason.

The Spinozan view thinks that this move away from epistemological voluntarism is a move in the right direction, but for the wrong reasons. For one thing, the Spinozan can actually grant people a decent degree of indirect epistemic control. If you know that you will

believe whatever you think, then you can control what you believe by controlling what you think. Although direct control over what we think is often quite difficult (Wegner 1994), indirect control can be had by the clever and sequestered. I think people intuitively pick up on this: it doesn't take much observation to see people who don't want to change their mind, so they bury their head in the sand. There's a reason why Fox News is consistently the highest rated news station of all time (more on this below).

The moral is even greater than just saying if you don't want to gain certain beliefs, then avert your eyes, for it seems that sub-personal cognitive systems often avoid unwanted information. Say you are the type of person who is committed to certain incompatible principles. This situation won't bother you if you don't notice that the principles are incompatible. Such blissful ignorance appears to be the norm. Imagine you are both anti-abortion and pro-death penalty. Prima facie, this is not the most compatible position to find oneself in. Of course, one can consistently hold both views without too much work. What's worth noting is that people often won't feel any tension between the views not because they think they are compatible, but because they fragment the beliefs in different belief systems. Dissonance theory gives us a way to understand why (and when) people do this: dissonant cognitions put people into dissonant states and dissonant states *hurt*. Consequently, people don't like being in them and, through conditioning, learn to avoid them. The lower one's threshold for dissonance the more one will avoid dissonance (Festinger 1957). And this avoidance comes in many forms. If you're a card carrying lefty with a low tolerance to dissonance, you'll stay away from AM talk radio. But you'll also be more apt to not see the consequences of your commitments. Most of us hold some contradictory commitments somewhere in our stock of beliefs. For those of us who with low tolerance to dissonance,

we're quite motivated avoid uncovering those contradictions. This process is a somewhat rational one—after all, there is a clear sense in which it is rational to stave off pain and the end result of these cases is keeping one's psychological immune system safe and secure.

So, thinking about how we can control our beliefs leads us to see one type of low grade irrationality. But dig a little further and you'll uncover weapons-grade irrationality. The Spinozan theory thinks that the move away from voluntarism is in general good, not just because we can't control our beliefs but because even our cognitive sub-systems, in the first instance at least, can't control what we believe—we just believe whatever we think. It is this datum that wrecks havoc with our concept of rationality and our view of what it is to be human.

I won't rehash the arguments from 2.2 in great detail, I'll just reiterate that the Spinozan theory rules out the ability to impartially doxastically deliberate in the first instance. Intuitively, a necessary condition on rationality is the ability to judiciously weigh evidence before taking that evidence in. To reiterate an earlier point, note that this is a different criticism from the ones at issue in the 'rationality wars': in those debates people's competence is generally not in doubt. Instead, what is debated is just how frequently people can utilize their basic competence. But the Spinozan attack on rationality goes after our basic competence, our ability to ever withhold assent.

The Spinozan theory creates the following dilemma for rationality: either the ability to impartially doxastically deliberate is not a precondition on rationality or people are (nomologically) necessarily irrational. Neither option is particularly appealing. Part of our concept of rationality is the ability to be a judicious cognizer; as philosophers we particularly pride ourselves on our ability to justify our beliefs and we have the expectation that these

justifications aren't just post-hoc rationalizations. However, if the Spinozan theory is right then we don't have the ability to deliberate about a proposition before believing the proposition.

That's just the start of the trouble for impartial deliberation. If the Spinozan theory is correct, not only would we be unable to withhold assent from propositions, but also we would be unable to impartially consider the beliefs that we do hold. Because of the confirmation bias we will have a quite partial deliberation strategy, one where we tend to search for confirming information while ignoring disconfirming information. Thus, at no point in our doxastic lives will we be able to consider propositions in a non-biased way. But it seems that our normative standards demand that a rational cognizer at least be able to impartially consider propositions at some point or other. So, the first horn is quite unappealing.

The second horn is also unpalatable. For years research has been mounting that people tend to be irrational in all sorts of domains: we ignore base rates, we're Dutch bookable, we have trouble working out probabilities, etc. However, all these cognitive illusions are set against a background presumption of rationality. We consider ourselves irrational in these ventures as compared to our normal rational conception of ourselves. The rational conception of ourselves is central to many theories of intentional ascription and linguistic communication (e.g. Davidson 2001, Dennett 1987). Take rationality away and it's hard to know what to make of so-called 'principles of charity.'

But on the other hand, if we are necessarily irrational, it's hard to make out what rationality amounts to. Pace whimsical sentiments about angels, we are supposed to be the

paradigmatic rational creature. If we were to give up our conception of ourselves as rational creatures, then it is unclear what the paradigm of rational creature would be.

I suppose one can get a handle on the dilemma if one has a theory of the relation between ought and can as they arise in epistemology. It seems plausible that there is a normative rule of epistemology that goes something like: 'don't believe a proposition unless you have some evidence for it.' If the Spinozan theory is right, then people can never follow this rule. If this rule is necessary to follow in order to be rational, then people can never be rational. But one might think that this situation is an untenable one, for rationality isn't a natural property—it's an evaluative term and its own that is supposed to apply to us. So maybe what we need to do is rethink our conception of rationality. On the other hand, if you think that it is fine to say that people ought to do what they can't, then perhaps you'd like to keep the rule stated above and insist on our rationality.

I feel the pull of both sides of the dilemma, thus I'm not sure which petard to foist myself upon. It does seem odd to have the concept of rationality not apply to us, but it strikes me as odder to separate belief fixation from rationality. We can concoct other conceptions of rationality that sidestep this problem, but I don't see how that'll make the dilemma problem away because it seems so plausible that believing and deliberating *in the right* way is necessary for rationality. Getting answers right is all fine and good, but how we come up with those answers is also important.

I raise this dilemma not to solve it, but only to point out that our concept of rationality is imperiled in a new way. If the Spinozan theory is correct, we will have to reconsider either our standards of rationality or our conceptions of ourselves. Perhaps a cherished metaphor will help drive home the Spinozan challenge to rationality. The Spinozan theory gives us

another way to understand the metaphor of Neurath's boat: we are always reconstructing our boat at sea because we never have any fixed point from which to adjudicate our beliefs. We're stuck with our beliefs, and even when we reject some, we are constantly drifting in the direction of the beliefs we hold, even if that direction is not particularly justifiable. We drift because our beliefs guide our searches towards confirming what we already believe, which in turn is in part a function of whatever we happen to have thought. And of course, the propositions we entertain are often a hodge-podge. Sometimes a thought pops in one's head not because of some computational inference process, but instead because of one's dinner choice. And presumably even the gourmands amongst us don't want to have our epistemology held hostage to our gustation.

5.2 False Histories and Bold Futures: Behaviorism, Cognitive Architecture, and Cognitive Science

In this section I will generate some different ideas about both where the Spinozan theory fits within the last 100 years or so of cognitive research and then I'll discuss a bit about the properties of the Spinozan theory that have been less than fully discussed in this essay. I'll end the section by saying a little about where I think future research in this area is likely to lead and give a glimpse at what I think the picture of the mind we end up with looks like.

5.2.1 Behaviorism and Beyond

There's a potted history of psychology that is widely proliferated: Introspectionists like Wundt and Titchener ushered in the beginnings of experimental psychology, but the behaviorists showed that their experimental procedure didn't have robust enough controls and so their results weren't repeatable or generalizable. The behaviorists then ushered in a

new era of objective psychology, but their spartan metaphysical scruples didn't allow them to ask the interesting questions. Hull showed his colleagues that they could all use a little C between their S and R and prince Chomsky and his minions showed that there could be a lot of structure in that C. Then we all lived happily ever after in a brave new interdisciplinary world, one where "friendship has become social cognition, affect is seen as a form of problem-solving, new-born perception
is subsumed under a set of transforming rules, and psychoanalysis is reread as a variant of information processing" (Kessen, 1981, p. 168). Supposedly one could also go to graduate school, study cognitive science, and then get a tenure track job.

But that story is at best misleading. Introspectionism wasn't all the rage across the academy and there were behavioristic ideas and criticisms floating around well before Watson got his hands anywhere near children. Dewey was already criticizing introspective methods a good thirty years before Watson commandeered the spotlight (for example, Dewey wrote about how experimentation now ''supplemented and corrected the old method of introspection'' (Dewey, 1884, p. 282; see also, Dewey 1896). Bretano had similar criticisms:

> If someone is in a state in which he wants to observe his own anger ranging within him, the anger must already be somewhat diminished, and so his original object of observation would have disappeared. The same impossibility is also present in all other cases. It is a universally valid psychological law that we can never focus our attention upon the object of inner perception (Brentano, 1874, p. 30; cited in Costall 2006).[111]

[111] Of course, Hume knew this too: "tis evident this reflection . . . would so disturb the operation of my natural principles as must render it impossible to form any just conclusion from the phenomenon. We must, therefore, glean up our experiments in this science from a cautious observation of human life, and take them as they appear in the common course of the world, by men's behaviour in company, in affairs, and in their pleasures" (Hume,1739-40 p. 46).

As mentioned earlier (footnote 85), even Wundt, the man christened as the father of Introspectionism, wasn't sold all that sold on it ('Introspective method relies either on arbitrary observations that go astray or on a withdrawal to a lonely sitting room where it becomes lost in self-absorption'', Wundt, 1900, p. 180, translated in Blumenthal, 2001, p. 125). In fact, Wundt was one of the harshest critics of the Introspectionist research conducted by two of his most famous students, Titchener and Kulpe. Furthermore, upon examining the research conducted in Wundt's laboratory it appears that he primarily relied on non-introspective methods, mainly time measurements and straightforward qualitative judgments of exogenously given stimuli. One psychological historian writes

> 'Introspection,' as such, was seldom used, and then in the following limited ways: (a) attempts to explain individual differences in the objective data, which was of course a matter of no systematic interest in Wundt's laboratory; (b) checks on the effectiveness of experimental manipulations, e.g. in regard to levels of attention (Costall 2006).

So, Watson didn't overturn a huge movement because there was no such movement. And Watson and Skinner weren't just overturned by Chomsky either, there were a lot of other figures and concurrently active schools of thought attacking the behaviorists, like the Gestalt psychologists, social psychologists, George Miller, Leon Festinger, etc. Finally, behaviorism didn't totally die. For example, Skinnerian semantics survived and evolved into contemporary versions of informational semantics a development that strikes me as progress, even if it's imperfect progress (Dretske 1981, Fodor 1987a). I'd now like to suggest that another strain of Behaviorism survives in the Spinozan theory.

The Spinozan theory continues with the behaviorist insight that the learning process, of which the belief forming process is presumably an instance, is not a rational process (i.e. one doesn't learn because of being compelled by reasonable evidence). The Spinozan theory

isn't the only post-Behaviorist theory to pick up on the fact that learning isn't wholly rational—indeed, developmental psychology is rife with stories of one-shot learning, fast mapping, and the like (see, e.g., Carey and Bartlett 1978). The irrationality inherent in learning is pointed out quite explicitly by Fodor when discussing modularity theory: "by definition modular processing means arriving at conclusions by attending to arbitrarily less than all of the evidence that is relevant and/or by considering arbitrarily fewer than all of the hypotheses that might reasonably be true" (Fodor 1987). What makes this type of learning irrational (or perhaps better, arational) is not that the system is ignoring lots of evidence and is not testing many hypotheses; rather, what is troubling is that the learning system ignores *relevant* evidence and *reasonable* hypotheses.

But the Behaviorists thought something a bit more radical than this, they believed that what was learned needn't bear *any* rational relation to the situation in which it was learned. One can see this in Pavlov when he writes

> Any natural phenomenon chosen at will may be converted into a conditioned stimulus…Any visual stimulus, any desired sound, any odor, and the stimulation of any part of the skin, whether by mechanical means or by the application of heat or cold… (Pavlov 1928).

It is this idea about the total disconnect between what is learned and what it is rational to have learned that the Spinozan theory picks up on. The Spinozan, like the Behaviorist, views belief fixation as essentially informationally encapsulated to the limit, which is another way of saying that the Spinozan interprets belief fixation as a reflex.

5.2.2 Belief Boxes, Leak-less Stores, Fragmented Belief Systems, and Future Research

A few more words about the picture of the mind that the Spinozan theory comports with seems appropriate before we end. The Spinozan theory sits uncomfortably with the popular and misleading metaphor of a 'belief box' (Schiffer 1987). Since everything that is tokened is believed, belief seems to be the default propositional attitude. The belief box metaphor was intended so that we can functionally individuate which propositions go in the box and which don't. But if we believe everything we think then we have no need to distinguish which propositions are in our cognitive store outside of the belief box, for if it's in the mind, it's in the box.

One might still care to posit, say, a desire box or a wish box, but I wouldn't suggest it. Although we might want a handy metaphor for how we individuate the states of affairs that we desire from the ones we don't, I don't see how the metaphor of a desire box would be all that illuminating. My hunch is that once we get a properly sanitized notion of desire, we will probably be equally surprised by what the properties of desire are. As mentioned, I'd like to think about this essay as an attempt at filling out some of the properties of belief. Functionalism about the propositional attitudes is a fairly widespread view and most (non-analytic) functionalists believe that scientific psychology will fill out the functional roles of belief. Here, we've focused most acutely on one of those functional roles: its intimate connection to entertaining. Although we have some reason to think that beliefs may be special, we just don't know how the other propositional attitude terms will pan out, and consequently, it is unclear whether talk of a 'desire box' and the like will be illuminating or stifling.

In terms of belief, I suggest that as opposed to discussing belief boxes, we focus of the fragmentation of belief. Every proposition we token may be believed, but that doesn't

entail that they are all housed in a single network of propositions, in a single 'web of belief' (Quine and Ullian 1978). As mentioned way back in chapter 2, it strikes me as intuitively plausible that one never loses beliefs for any cognitive reasons. Losing a belief is never a purely rational process. That is, even if you clearly see the falsity of your belief that P, you still can't just stop believing that P. I have a few reasons for finding this to be plausible—for one, we can plainly see that oftentimes we know something but can't recall it; that is, we know our store of beliefs often outstrips our search function. In other words it strikes me (as I think it struck Freud, see Freud 1920) as fruitful to think that memory itself is 'perfect', but our search function is far less than perfect. Of course the view isn't that one can't misremember or make up memories; rather, it's that beliefs stick in one's memory though finding them might be quite arduous (this claim will be put less mysteriously in the next paragraph). Some suggestive evidence for the view that memory is perfect in this sense comes from Mitchell (2006) where it is demonstrated that priming effects lasted on subjects 17 years after the original prime (!).

So, I agree with Freud and will assume that memory is more or less a perfect system. When I say that memory is 'perfect' what I'm proposing isn't that one never loses any beliefs; rather, it's that one can't lose any beliefs for a cognitive (or if you prefer, rational) reason (for example, one can't lose a belief because they see the falsity in the belief). More specifically, I don't think anyone ever loses beliefs for any *computational* reasons. I don't think there is ever a chain of reasoning that starts from premises about the dubiousness of a belief one holds and then ends with the belief being 'dropped' (which is perhaps part of the puzzle for why it's so damn hard to convince anyone to change their mind through rational argument).

People can of course lose beliefs in many ways. However, I don't suspect that any of these ways will be cognitive. You may lose a belief because of neural atrophy as you age; you may lose a belief because of spending ninety intimate minutes in a closet with a tank of nitrous oxide; you may lose a belief because you didn't eat enough choline when you were ten; you may lose a belief because your skull has just made friends with the pavement... All these ways of losing a belief are, to put it crudely, at least one level down from the cognitive level.[112] All in all, this is the way it should be. When discussing special sciences, we generally have to go one level down to discuss implementation and breakdown, so why should memory be any different?

[112] I don't know that there is *one* such cognitive level. My bet is that there are multiple levels that have been described as 'cognitive.' Let me give you an example to explain my unease. At the end of Fodor (1975), he discusses the limits of cognitive psychology. He writes, "It is, I think, the next thing to dead certain that some of the propositional attitudes we entertain aren't the result of computations. That isn't of course to say they aren't caused; it's just to say that their causes aren't psychological…Some mental states are, as it were, the consequence of brute incursions from the physiological level.; if it *was* the oysters that one ate that were to blame, then there will be no *computational* interpretation of the causal chain that leads from them to one's present sense that things could, on the whole, be better" (p200). Fodor's picture here is that cognitive psychology, and thus the cognitive level, is to be contrasted with psychophysics. He writes, "Psychophysical truths express the lawful contingency of events *under psychological description* upon events *under physical description;* whereas the truths of cognitive psychology express the computational contingencies of events which are homogenously (psychologically) described" (201). The idea is that the cognitive processes will all be rule-governed processes. However, the problems for this view come out a bit later when Fodor writes, "Cognitive psychology is about how rationality is structured, viz., how mental stated are contingent on each other." The rationality bug seems to have bitten Fodor—he, like many, many other philosophers of mind (see, Dennett, Dan) think that the hallmark of cognitive psychology is based in understanding rational relations amongst representations—relations that hold in virtue of the content of the mental states. However, there seems to be a thriving branch (or branches?) of psychology that studies 'irrational' relations amongst mental states (e.g. associative strengths between ideas, common heuristics and biases, associative relations between ideas and situations…). Are these generalizations to be put at the cognitive level? Presumably not for Fodor because the relations don't hold in virtue of the rational relations between content (associations may hold for semantic, but not rational, reasons [of course they may not too—one may associate mall lighting with vomit, for example]). Are these 'irrational' generalizations to be pitched at the level of psychophysics? That can't be so, for they don't have any physical description in them (Weber's Law explains lots of things, but it doesn't explain why DOG thoughts lead to DOG FOOD thoughts). My bet is that social psychology occupies a place in between the 'cognitive level' (although social psychological effects are clearly cognitive) and the 'psychophysical 'level. Perhaps Fodor has missed this because he doesn't believe in social psychology. For a view that social psychology is the "psychology of experience" (and hence, I suppose, in between the 'cognitive' and psychophysical levels of explanation) see Wegner and Gilbert (2000).

So now, imagine, at least for the sake of argument, that everything we entertain we believe and that we can never lose beliefs, just toggle their strengths. How could we set up such a mind so that we could manageably get by in the world? One avenue that I find appealing follows Lewis's suggestion that belief systems are fragmented (Lewis 1982). People hold contradictory beliefs but not all of these beliefs are held in the same cognitive sub-system. If we do have Spinozan minds, then delineating different beliefs systems will be very important for the overall functioning of the organism. For example, as Egan (ibid) argues based on reasons stemming from the 'paradox of the preface' (Makinson 1965), it is well-known that one cannot believe all the deliverances of a system and believe that the system is less than perfectly reliable. Now consider a Spinozan system. The Spinozan system can't help but believe what it tokens, so it will believe (e.g.,) all the deliverances of its perceptual systems. Now imagine that we have a person whose visual system is less than perfectly reliable. Such a person couldn't consistently believe all the deliverances of the system and believe that the system is less than perfectly reliable. Egan argues that because the system will always deliver many more beliefs not about the system (e.g., THERE'S A ROCK, THERE'S A BIRD…) than beliefs about the reliability of the system, the person will be forced, if she's shooting for consistency, to believe that her visual system is reliable and ignore evidence that speaks against this.

This is a difficult situation: if we have an unreliable belief forming mechanism, like a non-veridical visual system with a Spinozan psychology, we'd like to find out about it, but the above considerations make it seem as if all the evidence we could find to show us that the system is unreliable would be trumped by the beliefs formed by the deliverances of that system. Then how can we find out about the unreliability of such a system? What one could

do is fragment one's belief network, keeping the beliefs stemming from the visual system in one place and the beliefs about the visual system in another.

In fact, it may be that the evaluative aspect of the Spinozan architecture that aims for consistency only makes sense in terms of fragmented belief networks. Instead of there being a single overall credence function, the Spinozan should take seriously the possibility that there are many different fragmented functions at play. In order to evaluate how a given proposition coheres with the rest of one's framework of beliefs often means evaluating a candidate proposition with other, competing propositions that are inconsistent with the candidate proposition. So, how does one even strive for consistency in such a framework? By fragmenting their belief system so that, as much as possible, the inconsistent beliefs don't interact with each other. Then one's separate stocks of beliefs can become active in different situations—with context disambiguating which system to use at what point.

If this is on the right track, then a reasonable research strategy would be to analyze and individuate these sub-systems. It would be nice to see, e.g., under what conditions they come about how they come about, what the upper bound on the number of such systems is, under what conditions the subsystems get merged, whether infants and children fragment in similar ways to adults etc. Unsurprisingly, I have little interesting to say about how such research should proceed, but questions like this will occupy some of my future focus.

However, that doesn't mean that I don't have a hunch about one property such systems will share: automaticity. The strain of psychological thought that the Spinozan shares with behaviorism, modularity theory, and most of social psychology is based on the lack of control one has over one's cognitive system. The unbearable automaticity of being isn't just about behavior, it also applies to belief acquisition. It would thus surprise me very

little if, however the facts about fragmented belief work out, they work out in such a way that we, at the person-level and sub-personally, have little control over how and when such systems arise and in what situations they are active. This isn't just to point out that there is no 'ghost in the machine;' rather, it's to point out that there is no single machine too—there are many machines all lacking ghosts.

5.3 Propaganda

The Spinozan theory can help to elucidate the efficacy and working of propaganda. The second step in the Spinozan model is the effortful step of rejecting a previously held proposition. The negation of belief is how the Spinozan theory models deliberative processes; for the Spinozan theory deliberative processes can occur, but only after a proposition is believed. Yet this deliberative process is quite fragile. If one is under any cognitive load, the process is apt to short-circuit. Therefore, one needs all their mental faculties in order to be appropriately skeptical.

This is a fact that philosophers have intuited for some time. When Descartes escapes to his den to contemplate which of his beliefs are clear and distinct he is escaping from the world as much as is possible. He's trying to stave off all extraneous distractions. The armchair is supposed to be a place where one can contemplate unimpeded by worldly distractions.

But the ubiquity of worldly distractions abound. Take, for example, one's cable news networks. Your basic screen will involve a picture with a talking head on it, a blurb below the talking head, and a 'crawl' underneath both (see the pictures below). According to the Spinozan model, one can have one's attention split and reflexively form new beliefs, but one cannot be cognitively taxed and reject a given proposition. With all the distractions on the

screen, one cannot help but be distracted. Moreover, since one is constantly parsing propositions while watching the news, one is much more apt to blithely agree with what one hears. It takes a hefty amount of focus to drown out the distractions.

The Spinozan view also explains why television commercials are so effective in reaching their audiences. Suppose one is told "Coke tastes great" while being visually entranced by a scantily clad member of the opposite sex or while watching unpleasant colors swirl on a television screen. In this case one is more apt to be distracted while hearing the commercial's pitch and thus one is less likely to have the cognitive resources available to reject the given proposition.

The effects hold for the classroom too. For example, it is clear that many students take on their professors' pet projects. When students are in class they are more often than not self-regulating their behavior because they are worried about how they are being perceived, thus they are more apt to believe what they hear without having the requisite cognitive energy to go back and reject their belief. Thus, they end up believing their professors theories (regardless how crackpot those theories are).[113]

In sum, the Spinozan theory can open up new ways of looking at an age old phenomenon—public control of private minds.

5.4 Possible Future Experiment

Before closing, I'll suggest one experiment I've been devising that would lend more strong, direct support for the Spinozan theory. In the learning phase of the experiment participants are given twenty sentences, ten of which are obviously true ('OT' hereafter, e.g., '2+2=4'; 'Bachelors are unmarried men'; 'Washington D.C. is the capital of the United

[113] All concepts are innate my foot I might add.

States'; 'Barack Obama is the president of the United States' etc.), and ten of which are obviously false ('OF' hereafter, e.g., 'Your mom is a giant toaster eating moth'; '2+2=5'; 'An alligator is a basketball', etc.).[114] During the learning phase participants are given as much time as they'd like to contemplate the veracity of the sentences. Once they have decided on the truth of the proposition, they will push a button cuing the next sentence. Presumably, the participants will judge the ten obviously true statements as obviously true and the ten obviously false sentences as false. During the testing phase, participants will see five OT and five OF sentences, along with ten new sentences, five true ('NT' hereafter) and five false ('NF" hereafter). Participants will be asked to respond as quickly as they can whether the sentence is true or false.

The Spinozan view predicts that response times to OF should be slower than response times to OT, that response times to NF should be slower than response times to NT, and, most strikingly, that response times to OF should be slower than response times to NT. This is because the theory implies that when one considers the truth of 'Angkor is the capital of North Carolina' one first believes the proposition before rejecting it. However, a trace of this belief should still exist. So, when recalling a OF, the Spinozan predicts a two stage process: a recall and the subsequent rejection stage. It seems natural for the Cartesian view to predict that response times to OF sentences should be faster than response times to (e.g.) NT ones because people would have already considered and rejected the old ones, so they should be quicker (via priming suppose) to recall the answer they've already formed than they would be in assessing the new proposition.

[114] For both parts of the experiment the order of the sentences will be randomized and word frequency and sentence length will have to be controlled for.

If an experiment like this would have the results mentioned here, I think we'd have pretty strong evidence for the Spinozan view. Of course, I think we have pretty strong evidence for it anyway, but there is no reason to sneeze at further data.

5.5 Summary, Merciful Summary

So, what has been shown? I've argued that there is a slew of evidence against the intuitive and ubiquitous Cartesian theory of belief fixation. In its stead, I have offered a Spinozan theory of belief fixation. The Spinozan theory is the best going theory we have of how beliefs are acquired. The theory finds support from a variety of sources. Additionally, it can help explain many disparate phenomena. Furthermore, if the theory holds, then what appeared to be quite disparate psychological phenomena are all explicable with one elegant theoretical posit: tokening is believing.

My goal has not been to argue that the theory is necessarily true, rather my aim has been the milder end of establishing that the theory is a respectable hypothesis about belief acquisition. And respectable hypotheses are what we need, for we have an overwhelming dearth of plausible theories of belief acquisition. The Spinozan theory is a fecund program, one with wide-ranging applications and one that can unify and explain quite disparate findings in psychology while diffusing some philosophical paradoxes. If you don't find the theory plausible upon first read, I recommend rereading the paper, preferably while in a bustling café.

References

Anderson, C., M. Lepper, and L. Ross. 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.

Audi, R. 2008. "The Ethics of Belief: Doxastic Self-Control and Intellectual Virtue." *Synthese*, 161, 3: 403-418.

Bargh, J., and Chartrand, T. 1999. "The Unbearable Automaticity of Being." *American Psychologist*, 54: 462-479.

Bem, D. J. 1970. *Beliefs, Attitudes, and Human Affairs.* Belmont, Calif.: Brooks/Cole.

Berson, R. 1983. "Capgras Syndrome." *American Journal of Psychiatry*, 140 (8): 969-978.

Blasko, D., and C. Connine. 1993. "Effects of Familiarity and Aptness on Metaphor Processing." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19 (2): 295–308.

Blumenthal, A. 2001. "A Wundt Primer: The Operating Characteristics of Consciousness." In R. W. Rieber & D. K. Robinson, eds., *Wilhelm Wundt in History: The Making of a Scientific Psychology*. New York: Kluwer Academic/Plenum Publishers.

Brentano, F. 1874. *Psychology from an Empirical Viewpoint*. London: Routledge.

Brigard, F., Mandelbaum, E., and Ripley, D. 2009. "Responsibility and the Brain Sciences." *Ethical Theory and Moral Practice*, 12 (5): 511-524.

Block, J. 1965. *The Challenge of Response Sets.* New York: Appleton Century Croft.

Cacioppo, J. T., and R. E. Petty. 1982. "The Need for Cognition." *Journal of Personality and Social Psychology* 42 (1): 116–31.

Carey, S. and Bartlett, E. 1978. "*Acquiring a Single New Word*." In *Papers and Reports on Child Language Developmen*t 15, 17-29.

Chapman, G. and Johnson, E. 2002. "Incorporating the Irrelevant: Anchors in Judgments of Belief and Value." In T. Gilovich, D. Kahneman, and A. Tversky, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Christodolou, G. 1977. “The Syndrome of Capgras.” *British Journal of Psychiatry*, 130: 556-564.

Clark, H. and Clark, E. 1977. *Psychology and Language*. New York: Harcourt, Brace, Jovanovich.

Cooper, J. 2007. *Cognitive Dissonance: 50 Years of a Classic Theory*. London: Sage Publications Ltd.

Costall, A. 2006. “Introspectionism and the Mythical Origins of Scientific Psychology.” *Consciousness and Cognition*, 15: 634-654.

Couch, A and Keniston, K. 1960. “Yea-Sayers and Nay-Sayers: Agreeing Response Set as a Personality Variable.” *Journal of Abnormal and Social Psychology*, 62: 175-179.

Crane, T. 2001. *Elements of Mind*. Oxford: Oxford University Press.

Davidson, D. 2001. “Radical Interpretation”, in *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Dehaene, S. 1997. *The Number Sense*. NY: Oxford University Press.

Descartes, R. 1988. *The Philosophical Writings of Descartes*. Cambridge: Cambridge University Press.

Dennett, D. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

———. 1991. “Real Patterns.” *Journal of Philosophy*, 88 (1): 27-51.

———. 1998. *Brainchildren*. Cambridge, Mass.: MIT Press.

Dewey, J. 1896. “The Concept of the Reflex Arc in Psychology.” *Psychological Review*, *3*: 357-370.

———. 1918. “Concerning Alleged Immediate Knowledge of Mind.” *Journal of Philosophy*, 15 (2): 29-35.

Doris, J. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*. Cambridge, Mass: MIT Press.

———. 1993. “Conscious Experience.” 102 (406): 263-283.

———. 2004. “Knowing What You Think vs. Knowing That You Think It” in Richard Schantz, ed., *The Externalist Challenge*. Berlin: De Gruyter.

———.Unpublished. “I Think I Think—Therefore I Am, I Think.”

Dunlap, K. 1912. “The Case Against Introspection.” *Psychological Review*, 19: 404-413.

Dutton, D., and A. Aron. 1974. “Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety.” *Journal of Personality and Social Psychology* 30 (4): 510–17.

Egan, A. 2008. “Seeing and Believing: Perception, Belief Formation, and the Divided Mind.” Philosophical Studies, 140 (1): 47-63.

Epley, N. 2004. “A Tale of Tuned Decks? Anchoring as Accessibility and Anchoring as Adjustment.” In D. J. Koehler and N. Harvey, eds., *The Blackwell Handbook of Judgment and Decision Making*. Oxford: Blackwell Publishers.

Epley, N., and T. Gilovich. 2001. “Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors.” *Psychological Science* 12 (5): 391–96.

———. 2006. “The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient.” *Psychological Science* 17 (4): 311-318.

Epley, N., B. Keysar, L. Van Boven, and T. Gilovich. 2004. “Perspective Taking as Egocentric Anchoring and Adjustment.” *Journal of Personality and Social Psychology* 87 (3): 327–39.

Evans, G. 1982. *Varieties of Reference*. New York: Oxford University Press.

Festinger, L. 1957. *Theory of Cognitive Dissonance*. Palo Alto: Stanford University Press.

Festinger, L., and N. Maccoby. 1964. “On Resistance to Persuasive Communication.” *Journal of Abnormal and Social Psychology* 68 (4): 359-66.

Fodor, J. 1968. *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York: Random House.

———. 1975. *The Language of Thought*. New York: Thomas Y. Crowell.

———. 1981a. “Propositional Attitudes.” In *RePresentations*: *Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Mass.: MIT Press.

———. 1981b. “The Present Status of the Innateness Controversy.” In *RePresentations*: *Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, Mass.: MIT Press.

———. 1983. *Modularity of Mind*. Cambridge, Mass.: MIT Press.

———. 1987. *Psychosemantics*. Cambridge, Mass: MIT Press.

———. 1987. “Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres.” In J. Garfield, ed., *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, Mass.: MIT Press.

———. 1998 *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

———. 2000. *The Mind Doesn’t Work That Way*. Cambridge, Mass.: MIT Press.

———. *LOT 2: The Language of Thought Revisited*. NY: Oxford University Press.

Ford, K., and Z. Pylyshyn, eds. 1996. *The Robot’s Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Norwood, N.J.: Greenwood Publishing Group.

Frazer, S., and Roberts, J. 1994. “Three Cases of Capgras Syndrome.” *British Journal of Psychiatry*, 164: 557-559.

Freud, S. 1920. *Introductory Lectures on Psychoanalysis*. New York: W. W. Norton and Company.

Frost, R. 1995. *Robert Frost: Collected Poems, Prose and Plays*. New York: Library of America.

Gallistel, C., and R. Gelman. 1992. “Preverbal and Verbal Counting and Computation.” *Cognition* 44 (1-2): 43-74.

Gendler, T. 2000. “The Puzzle of Imaginative Resistance.” *Journal of Philosophy* 97 (2): 55–81.

———. 2008. “Alief and Belief.” *Journal of Philosophy* 105 (10): 634-63.

———. 2008. “Alief in Action (and Reaction).” *Mind and Language* 23 (5): 552–85.

Georgalis, N. 2006. *The Primacy of the Subjective*. Cambridge: MIT Press.

Gigerenzer, G. 2007. *Gut Feelings: The Intelligence of the Unconscious*. London: Viking.

Gilbert, D. 1991. “How Mental Systems Believe.” *American Psychologist* 46 (2): 107-19.

———.1993. The Ascent of Man: Mental Representation and the Control of Belief.” In D. Wegner and J. Pennebaker, eds., *The Handbook of Mental Control*. Englewood Cliffs, N.J.: Prentice-Hall.

———. 2002. “Inferential Correction.” In T. Gilovich, D. Kahneman, and A. Tversky, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Gilbert, D., D. Krull, and M. Malone. 1990. “Unbelieving the Unbelievable: Some Problems in the Rejection of False Information.” *Journal of Personality and Social Psychology* 59 (4): 601-13.

Gilbert, D., R. Tafarodi, and P. Malone. 1993. “You Can’t Not Believe Everything You Read.” *Journal of Personality and Social Psychology* 65 (2): 221-33.

Goldman, A. 1993. “The Psychology of Folk Psychology.” *Behavioral and Brain Sciences*, 16: 15-28.

Hasson,U., Simmons, J., and Todorov, A. 2005. “Believe It or Not: On the Possibility of Suspending Belief.” *Psychological Science*, 16 (7): 566-571.

Hasson, U., and S. Glucksberg. 2006. “Does Negation Entail Affirmation? The Case of Negated Metaphors.” *Journal of Pragmatics* 38: 1015–32.

Heider, F. 1944. “Social Perception and Phenomenal Causality.” *Psychological Review*, 51: 358-374.

Hirstein, W., and Ramachandran, V. 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*, 264: 437-444.

Huebner, B. 2009. “Troubles with Stereotypes for Our Spinozan Psychology.” *Philosophy of Social Science* 39 (1): 63–92.

Hume, D. 1739-40. *A Treatise of Human Nature*. London: Penguin Books.

Isen, A. and Levin, P. 1972. “The Effect of Feeling Good on Helping: Cookies and Kindness.” *Journal of Personality and Social Psychology*, 21 (3): 384-388.

Jacowitz, K., and D. Kahneman. 1995. “Measures of Anchoring in Estimation Tasks.” *Personality and Social Psychology Bulletin* 21 (11): 1161–66.

James, W. 1896/1992. *The Will to Believe.* New York :Library of America.

Jones, E. 1979. “The Rocky Road from Acts to Dispositions.” *American Psychologist* 34 (2): 107–17.

Kahneman, D., and A. Tversky. 1974. “Judgment under Uncertainty: Heuristics and Biases.” *Science* 185 (4157): 1124–31.

———. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science* 211 (4481): 453–58.

Kessen, W. 1981. “Early Settlements in New Cognition.” *Cognition,* 10: 167-171.

Keysar, B. 1993. “Common Sense and Adult Theory of Communication.” *Behavioral and Brain Sciences* 16: 54.

———. 1994. “The Illusory Transparency of Intention: Linguistic Perceptive Taking in Text.” *Cognitive Psychology,* 26: 165-208.

Keysar, B., and Barr, D. 2002. “Self Anchoring in Conversation: Why Language Users Do Not Do What They ‘Should’.” In T. Gilovich, D. W. Griffin, & D. Kahneman, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*. (pp. 150-166). Cambridge University Press.

Klayman, J., and Y. Ha. 1987. “Confirmation, Disconfirmation, and Information in Hypothesis Testing.” *Psychological Review* 94(2): 211–28.

Knowles, E., and C. Condon. 1999. “Why People Say ‘Yes’: A Dual Process Theory of Acquiescence.” *Journal of Personality and Social Psychology* 77 (2): 379–86.

Kohler, W. 1929. *Gestalt Psychology*. New York: Liveright.

Kruger, J. 1999. “Lake Wobegon Be Gone!: The ‘Below Average Effect’ and the Egocentric Nature of Comparative Ability Judgments.” *Journal of Personality and Social Psychology* 77 (2): 1121–34.

Kunda, Z., G. Fong, R. Sanitoso, and E. Reber. 1993. “Directional Questions Direct Self-Conceptions.” *Journal of Experimental Social Psychology* 29 (1): 63–86.

Langer, E. 1975. “The Illusion of Control.” *Journal of Personality and Social Psychology,* 32, 2: 311-328.

Larson C., and Sullivan, J. 2006. “Watson’s Relation to Titchener.” *Journal of the History of the Behavioral Sciences,* 1 (4): 388-354.

Lewis, D. 1982. “Logic for Equivocators.” *Nous,* 16, 431-444.

Long, A., and D., Sedley, eds. 1987. *The Hellenistic Philosophers: Translations of the Principal Sources with Philosophical Commentary, Vol. 1*. Cambridge: Cambridge University Press.

Lormand, E. 1996. “Nonphenomenal Consciousness.” *Noûs,* 30: 242-61.

Lycan, W. 1986. “Tacit Beliefs.” In R. Bogdan, ed., *Belief*. Oxford: Oxford University Press.

———. 1987. *Consciousness.* Cambridge: MIT Press.

———. 1996. *Consciousness and Experience.* Cambridge, Mass: MIT Press.

———. Forthcoming. “Phenomenal Intentionalities.” *American Philosophical Quarterly*.

Maisson, J. and Dufosse, M. 1988. “Coordination Between Posture and Movement: Why and How?” *News in Physiological Sciences*, 3: 88-93.

Makinson, D. 1965. “Paradox of the Preface.” *Analysis* 25: 205-207.

Mandelbaum, E., and Ripley, D. Unpublished Manuscript. “Explaining the Abstract/Concrete Paradoxes in Moral Psychology: NBAR Theory.”

McDowell, J. 2009. *The Engaged Intellect: Philosophical Essays*. Cambirdge: Harvard University Press.

McGee, R. 1967. “Response Set in Relation to Personality: An Orientation.” In I. A. Berg, ed., *Response Set in Personality Assessment* (pp. 1-31). Chicago: Aldine.

Meck, W., and Church, R. 1983. “A Mode Control Model of Counting and Timing Processes.” *Journal of Experimental Psychology: Animal Behavior Processes*: 9 (3): 320-334.

Milgram, S. 1974. *Obedience to Authority*. New York: Harper Perennial.

Mitchell, D. 2006. “Non-Conscious Priming after 17 Years: Invulnerable Implicit Memory?” *Psychological Science* 17 (11): 925–29.

Moyer, R., and Landauer, T. 1967. “Time Required for Judgments of Numerical Equality”. *Nature,* 215 (5109): 1519-1520.

Mussweiler, T., and F. Strack. 1999. “Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model.” *Journal of Experimental Social Psychology* 35 (2): 136–64.

———. 2000. “Numeric Judgment under Uncertainty: The Role of Knowledge in Anchoring.” *Journal of Experimental Social Psychology* 36 (5): 495–518.

Nisbett, R., and L. Ross. 1980. *Human Inference: Strategies and Shortcomings of Human Inference*. Englewood Cliffs, N.J.: Prentice Hall.

Nisbett, R., and T. Wilson. 1977. “Telling More than We Can Know: Verbal Reports on Mental Processes.” *Psychological Review* 84 (3): 231–59.

Pavlov, I. 1928. *Lectures on Conditioned Reflexes. Twenty-Five Years of Objective Study of the Higher Nervous Activity of Animals*. Trans. by W. H. Gantt & G. Volborth. New York: International Publishers.

Pitt, D. 2004. “The Phenomenology of Cognition Or *What Is It Like to Think that P?*,” *Philosophy and Phenomenological Research*, 69: 1-36.

Pollock, J. 1986. *Contemporary Theories of Knowledge*. Totowa, NJ: Rowan and Littlefield.

Price, H. 1990. “Why ‘Not.’” *Mind* 99 (394): 221–38.

Priest, G. 2006. *Doubt Truth to be a Liar*. New York: Oxford University Press.

Prinz, J. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. New York: Oxford University Press.

Prinz, J. 2004. “The Fractionation of Introspection.” *Journal of Consciousness Studies*, 11 (7-8): 40-57.

Pryor, J. 2000. “The Skeptic and the Dogmatist.” *Nous* 44 (4): 517–49.

Pylyshyn, Z. 1989a. “Computing in Cognitive Science.” In M. I. Posner, ed., *Foundations of Cognitive Science*. Cambridge, Mass.: MIT Press.

———. 1989b. “The Role of Location Indexes in Spatial Perception: A Sketch of the FINST Spatial-Index Model.” *Cognition*, 32: 65-97

Quattrone, G. 1982. “Overattribution and Unit Formation: When Behavior Engulfs the Person.” *Journal of Personality and Social Psychology*, 42 (4): 593-607.

Quine, W., and Ullian, J. 1978. *The Web of Belief*. New York: Random House.

Ramerick, S. 2002. “Automated Choice Heuristics.” In T. Gilovich, D. Kahneman, and A. Tversky, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press.

Recanati, F. 1997. “Can We Believe What We Do Not Understand?” *Mind and Language*, 12 (1): 84-100.

Ripley, D. Unpublished Manuscript. “Contradictions at the Borders.”

Rodin, J. and Langer, E. 1977. “Long Term Effects of a Control Relevant Intervention with the Institutionally Aged.” *Journal of Personality and Social Psychology*, 35, 12: 897-902.

Roedinger, H., and McDermott, K. 1995. “Creating False Memories: Remembering Words not Presented in Lists.” *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21 (4): 803–14.

Rosen, G. 2004. “Skepticism about Moral Responsibility.” *Philosophical Perspectives, Ethics,* 18: 295-313.

Roskies, A. 1999. “The Binding Problem.” *Neuron*, 24 (1): 7-9.

Rozin, P. Markwith, M. and Ross, B. 1990 “The Sympathetic Magical Law of Similarity, Nominal Realism, and Neglect of Negatives in Response to Negative Labels,” *Psychological Science,* 1, (6): 383-384.

Ross, L., M. Lepper, and M. Hubbard. 1975. “Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm.” *Journal of Personality and Social Psychology* 32 (5): 880–92.

Ross, L. 1977. “The Intuitive Psychologist and His Shortcomings: Distortions in the Attribution Process.” In L. Berkowitz, ed., *Advances in experimental social psychology* (v. 10). New York: Academic Press.

Ryle, G. 1949. *The Concept of Mind*. Chicago: University of Chicago Press.

Samuels, R., S. Stich, and M. Bishop. 2002. “Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear.” In R. Elio, ed., *Common Sense, Reasoning and Rationality*. New York: Oxford University Press.

Schacter, D., K. Norman, and W. Koutstaal. 1997. “The Recovered Memories Debate: A Cognitive Neuroscience Perspective.” In C. Martin, ed., *Recovered Memories and False Memories: Debates in Psychology*. New York: Oxford University Press.

Schachter, D., Norman, K., & Koutstaal, W. (1998), ‘The Cognitive Neuroscience of Constructive Memory.’ *Annual Review of Psychology*, 49: 289–318.

Schiffer, S. 1987. *Remnants of Meaning*. Cambridge, Mass: MIT Press.

Sherman, J., and G. Bessenoff. 1999. “Stereotypes as Source-Monitoring Cues: On the Interaction between Episodic and Semantic Memory.” *Psychological Science* 10 (2): 106–10.

Sherman, D. K., and G. L. Cohen. 2006. “The Psychology of Self-Defense: Self-Affirmation Theory.” In M. P. Zanna, ed., *Advances in Experimental Social Psychology* 38. San Diego: Academic Press.

Skinner, B. 2002. *Beyond Freedom and Dignity*. Indianapolis: Hackett Publishing Company.

Slobin, J. 1966. “Grammatical Transformation and Sentence Comprehension in Childhood and Adulthood.” *Journal of Verbal Learning and Verbal Behavior*, 5, 219-227.

Snyder, M., and Swann, W. 1978. “Hypothesis Testing in Social Interaction.” *Journal of Personality and Social Psychology*, 36: 1202-1212.

Snyder, M., and Campbell, B. 1980. “Testing Hypotheses about Other People: The Role of the Hypothesis.” *Personality and Social Psychology Bulletin*, 6: 421-426.

Sperber, D. 1985. “Apparently Irrational Beliefs.” In *On Anthropological Knowledge: Three Essays*. Cambridge: Cambridge University Press.

Storms, M., and Nisbett, R. “Insomnia and the Attribution Process.” Journal of Personality and Social Psychology, 16 (2): 319-328.

Swann, W., Giuliano, T., and Wegner, D. 1982. “Where Leading Questions Can Lead: The Power of Conjecture in Social Interaction.” *Journal of Personality and Social Psychology*, 42: 1025-1035.

Spinoza, B. 1677. *Ethics*. Indianapolis: Hackett.

Strack, F., and T. Mussweiler. 1997. “Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility.” *Journal of Personality and Social Psychology* 73 (3): 437–46.

Steele, C. M. 1988. “The Psychology of Self-Affirmation: Sustaining the Integrity of the Self.” In L. Berkowitz, ed., *Advances in Experimental Social Psychology* 21. San Diego: Academic Press.

Stich, S. 1978. Beliefs and Subdoxastic States, *Philosophy of Science*, 45: 499-518.

Trott, D., and Jackson, D. 1967. “An Experimental Analysis of Acquiescence. *Journal of Experimental Research in Personality*, *2*: 278 -288.

Tulving, E. 1983. Elements of Episodic Memory. New York: Oxford University Press.

Tversky, A. 1972. “Elimination by Aspects: A Theory of Choice.” *Psychological Review*, 79: 281-299

Tversky, A. and Griffin, D. 1992. “The Weighing of Evidence and the Determinants of Confidence.” *Cognitive Psychology*, 24: 411-435.

Walton, K. 1978. “Fearing Fictions.” *Journal of Philosophy* 75 (1): 5–27.

Washburn, M. 1922. “Introspection as an Objective Method.” *Psychological Review*, 29: 89-112.

Wason, P. 1961. “Responses to Affirmative and Negative Binary Statements.” *British Journal of Psychology*, 52: 133-142.

Wason, P., and P. Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, Mass.: Harvard University Press.

Watson, J. 1913. “Psychology as the Behaviorist Views It.” *Psychological Review*, 20, 158-177.

Wegner, D. 1984. “Innuendo and Damage to Reputation.” In *Advances in Consumer Research* 11: 694-696.

———. *White Bears and other Unwanted Thoughts: Suppression, Obsession, and the Psychology of Mental Control*. New York: The Guilford Press.

———. 2002. *The Illusion of Conscious Will*. Cambridge: MIT Press.

Wegner, D., Wenzlaff, R., Kerker, R., and Beattie, A. (1981). “Incrimination Through Innuendo: Can Media Question Become Public Answers?” *Journal of Personality and Social Psychology* 40: 822-832.

Wegner, D., G. Coulton, and R. Wenzloff. 1985. “The Transparency of Denial: Briefing in the Debriefing Paradigm.” *Journal of Personality and Social Psychology* 49 (2): 338–46.

Wegner, D. and Gilbert, D. 2000. “Social Psychology: The Science of Human Experience.” In H. Bless and J. Forgas, eds., *Subjective Experience in Social Cognition and Behavior*. Philadelphia, PA: Psychology Press.

Williamson, T. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wilson, T. D., Lindsey, S., & Schooler, T. 2000. “A Model of Dual Attitudes.” *Psychological Review*, 107: 101-126.

Wood, J., Perunovic, W., and Lee, J. 2009. “Positive Self-Statements: Power for Some, Peril for Others.” *Psychological Science, 20,* 860-866.

Zanna, M. P., and J. Cooper. 1974. “Dissonance and the Pill: An Attribution Approach to Studying the Arousal Properties of Dissonance.” *Journal of Personality and Social Psychology* 29 (5): 703–09.