Inquiry: An Interdisciplinary Journal of Philosophy
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/sinq20

Thinking is Believing

Eric Mandelbaum[a]
[a] Harvard University, USA
Published online: 16 Dec 2013.

To cite this article: Eric Mandelbaum (2014) Thinking is Believing, Inquiry: An Interdisciplinary Journal of Philosophy, 57:1, 55-96, DOI: 10.1080/0020174X.2014.858417

To link to this article: http://dx.doi.org/10.1080/0020174X.2014.858417

Inquiry, 2014
Vol. 57, No. 1, 55–96, http://dx.doi.org/10.1080/0020174X.2014.858417

Thinking is Believing

ERIC MANDELBAUM

Harvard University, USA

(Received 13 November 2013)

ABSTRACT *The idea that people can entertain propositions without believing them is widespread, intuitive, and most probably false. The main goal of this essay is to argue against the claim that people can entertain a proposition without believing it. Evidence is presented demonstrating that we cannot withhold assent from any proposition we happen to consider. A model of belief fixation is then sketched and used to explain hitherto disparate, recalcitrant and somewhat mysterious psychological phenomena. The proposed model is one where beliefs are the automatic output of a computationally null belief acquisition reflex. In short, the model holds that the mere activation of a mentally represented truth apt proposition leads to immediately believing it. The essay concludes by considering some consequences that the proposed model of belief acquisition has for our concept of rationality.*

'Just keep moving forward and faith will come to you.'[1]

I. Belief Fixation, Doxastic Deliberation, and Rationality

Suppose that you have just stubbed your toe on a rock. If you are like some people, you will, at least momentarily, be angry *at the rock*. Even though you might know that the rock is not an appropriate recipient of your reactive attitude, often enough you cannot help but be angry at it.

Although we frequently feel emotions that are rationally groundless, we tend to assume that this is not equally true of our beliefs. If I ask you to please not believe what I am about to say (because, e.g., I am merely parroting someone else's falsehood), it seems plausible that you will be able to not believe what I am about to say. If I tell you that I am 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

Correspondence Address: Eric Mandelbaum, Department of Philosophy, Emerson Hall, Harvard University, 25 Quincy Street, Cambridge, MA 02138, USA. Email: eric.mandelbaum@gmail.com

[1]Wallace, *Everything and More*, 148.

the way that you may, for example, automatically get excited when hearing of a rare and tantalizing opportunity.

In what follows I argue that this plausible assumption is false. Just as we get angry at the rock while knowing full well that it is not an appropriate object of our anger, so too we generally believe what people say even when we know that what they are saying is false.[2] Like the elicitation of many emotions, belief formation is initially insensitive to our background beliefs.[3]

***

The idea that we can contemplate a proposition without believing it has been accepted in philosophy since at least the time of the Stoics,[4] and remains widespread in contemporary debates concerning everything from modularity theory to epistemology. To take just one example, Jerry Fodor writes:

> 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 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.[5]

Note that this story assumes that our central systems examine how different 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 conservative, gradual process that (ideally) takes into account all the relevant data in one's information store before assenting to any proposition.[6] Fodor's view is quite indicative of the field at large. Belief fixation is hypothesized to be a slow, conservative process, in part, to 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 in which propositions are automatically and reflexively believed, simply by being entertained. Discovering that belief fixation regularly avoids any interaction with background

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

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

[4]Long and Sedley, *Hellenistic Philosophers*, 438–61.

[5]Fodor, *Modularity of Mind*, 102.

[6]Hence Fodor writes things like 'the fixation of perceptual belief is the evaluation of such hypotheses in light of the *totality of background theory*'. Fodor, 'Observation Reconsidered', 24 (emphasis added).

information would be a very interesting and surprising fact about the mind. It would call into question the status of the idea that beliefs are fundamentally products of rational agency.[7]

The consequences of such a radical departure from the standard view extend beyond the topic of belief fixation. The ability to withhold assent from propositions that we entertain is a crucial part of our picture of an important variety of doxastic deliberation: the ability to consider propositions while suspending judgment. When first encountering a proposition, we take ourselves to be able to consider it while remaining neutral as to its truth. If we found creatures that regularly could not help but believe whatever idea they happen to entertain, we would be inclined to regard them as continually engaging in an irrational practice. Sadly, we seem to be such creatures.

A critique of rationality stemming from our inability to deliberate impartially differs from the contemporary 'rationality wars' criticisms.[8] Recent decades have brought heated debates over how rational people are, but these debates cluster around whether people tend to answer some particular problem correctly. One need not look hard to find claims that people are irrational because they, for example, fall prey to cognitive illusions, use fast and frugal heuristics, let emotions dictate their moral reasoning, and so forth. Throughout these debates, a cornerstone of our rationality has remained beyond critique: our ability to entertain propositions without believing them.[9] This ability has received scant attention and has endured very 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 shall have to reconsider the nature of doxastic deliberation. This is because, if the proposed theory is correct, then truly impartial doxastic deliberation is impossible—we would never begin doxastic deliberation from a neutral starting point. Instead, the best we could hope for is a type of deliberation that can undo beliefs we have acquired for free merely by entertaining certain propositions. Consequently, the theory of belief fixation defended here 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.

But before we get there, let us first be clear about the notion of belief with which we will be working. The notion of belief that is operative throughout

[7]A note about the scope of the claim: the arguments for the vast majority of the essay cover initial belief fixation, but remain neutral as to what capacities are exercised when one reconsiders a proposition already believed, though this topic briefly arises in the final section.

[8]Samuels, Stich, and Bishop, 'Ending the Rationality Wars'.

[9]For example, when arguing over whether the use of heuristics is ecologically rational, all parties assume that the information that heuristics process can initially be rejected. It is only after the acceptance of information that the question of the efficacy of our information processing techniques arises.

this paper will be the quotidian one 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.[10] Rather, belief will be understood similarly to the way one understands credences.[11]

Additionally, the essay assumes a certain amount of 'realism' about belief. This project is of a piece with the search for a state in cognitive science that shares a certain 'spiritual similarity'[12] with the folk-psychological notion of beliefs. As such, the essay works toward building a psychofunctional theory of beliefs. But one need not buy into any psychofunctional theory to assess the arguments in the paper. For the present purposes, all that one needs to assume is that beliefs are causally efficacious states—ones that, for example, are caused by perception, serve as the premise of inferences, and interact with desire to cause behavior. Not much else is assumed about belief, at least nothing that will affect the arguments that follow.

In what follows, I compare two theories of belief fixation. Ultimately, I argue that one of these theories is false and that the other theory can unify and explain a plethora of seemingly disparate phenomena and so should be taken quite seriously. But for now, let us consider a ubiquitous and influential theory of cognitive architecture: the Cartesian theory of belief fixation.

II. 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).[13] But it is worth asking: when Descartes was sitting beside the fire contemplating which propositions to believe in, what was he

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

[11]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 (in combination with the appropriate desires). Presumably a belief with a credence of 0.0001 will not produce any behavior; on the other hand, a belief need not have a credence of 0.9 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 propositions causes one's credences to go at least that high. This said, there will be little talk of credences in what follows, for my preferred analysis of what credences are is cashed out in terms of resistance to disconfirming evidence and not something akin to betting procedures.

[12]Fodor, *Psychosemantics*.

[13]Descartes, *Philosophical Writings*.

actually trying to do? He was attempting to first *entertain* an idea, then *contemplate* its truth, and finally *decide* what to assent to and what to *withhold* belief from. Descartes's attempt presupposed a serial model of belief fixation, according to which one first entertains a proposition, then subsequently either believes the proposition or withholds assent from the proposition.[14] 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 (following Gilbert) will term 'the Cartesian theory of belief fixation'.

What I am here calling the 'Cartesian Theory' consists of the following claims:

(1) People have the ability to entertain propositions that arise in the mind, whether through perception or imagination, before accepting those propositions.
(2) Accepting and rejecting a proposition use the same mental processes, and consequently, should be affected by performance constraints in similar ways.[15] I will sometimes refer to the Cartesian position as a

[14]Although this scenario admittedly paints Descartes with a broad brush, some relevant literature has interpreted Descartes as attempting the project I sketch out. See, for example, Gilbert, 'How Mental Systems Believe'; Huebner, 'Troubles with Stereotypes'. Nevertheless, there are some reasons to believe that Descartes actually was not a Cartesian in this sense. Some historians like Alan Nelson (personal communication) interpret Descartes's epistemic methodology as such: assume Descartes wants to assess the truth of the proposition that Santa Claus exists. Call this proposition S. Descartes's 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 shall ignore this element, which strikes me as regress prone.) The next step is to think of situations that 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 Descartes does not believe one can just think of negation as such. Nelson's Descartes holds a variation on the view that I am promoting; he believes that people believe everything they think because they do not have the ability to withhold assent. Rather, what people can do instead 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 superweak if it leads to a different belief that would entail the falsity of S (ironically, if this reading is right then my analysis of credences is very similar to Descartes's). A strong belief is a belief that does not automatically lead to a second belief, which destroys our consciousness of the first belief. So perhaps Descartes was not a Cartesian in the sense expressed in the main text. That does not really matter because an overwhelming majority of contemporary philosophers and cognitive scientists are. If one would like, they can substitute Pollock or Fodor (or anyone else who has the modular/central systems distinction) in for Descartes. See Pollock, *Contemporary Theories of Knowledge*; Fodor, *Concepts, Language of Thought, Modularity of Mind*.

[15]I use the phrases 'accepting a proposition' and 'believing a proposition' interchangeably; likewise for 'rejecting a proposition' and 'disbelieving a proposition'. No doubt, in the present climate doing so is controversial, but I do not have the space to argue for such usage. The idea

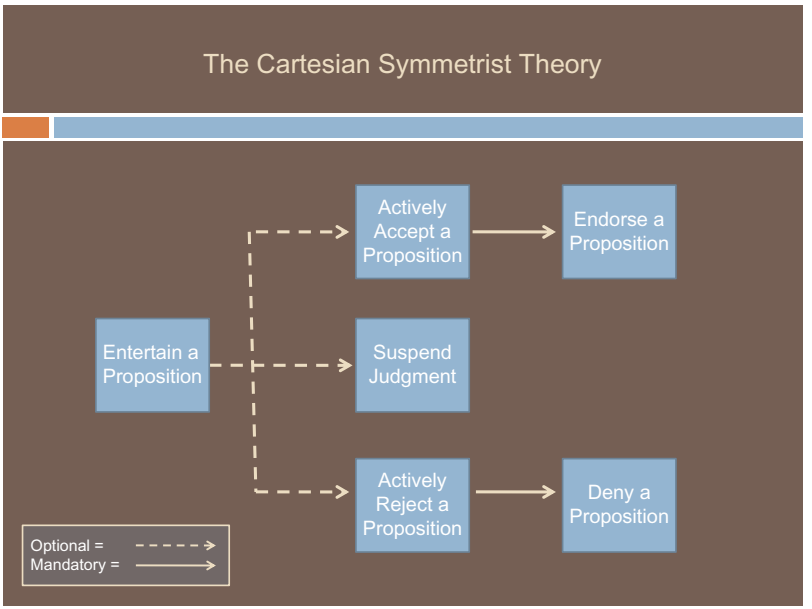

Figure 1. The Cartesian symmetrist theory.

Note: The dotted lines represent optional links, and solid lines necessary links.

'symmetrist' position, because it treats accepting and rejecting symmetrically.

(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.[16]

The Cartesian theory is intuitive, widely accepted, and rarely, if ever, argued for. It is assumed throughout many areas of both philosophy and psychology.[17] Moreover, the serial view of belief fixation that the theory

that the states discussed are indeed beliefs and not some form of 'mere acceptances' (and in fact that the theory proposed here can serve as a model for a full-fledged psychofunctional theory of belief) is discussed in Section IV.iv. See, for example: Velleman, 'On the Aim of Belief'; Stalnaker, *Inquiry*; Cohen, *Essay on Belief and Acceptance*; Bratman, 'Practical Reasoning'; Tuomela, 'Belief vs. Acceptance'; van Fraasen, *Scientific Image*.

[16]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). According to my view, even a fast-moving brick cannot derail one's passive assent.

[17]For example: Quine, *Word and Object*; Milgram, *Obedience to Authoriry*; Fodor, *Concepts, Language of Thought, Modularity of Mind*; Dennett, *Intentional Stance*; Pylyshyn, 'Computing in Cognitive Science'; Ford and Pylyshyn, *Robot's Dilemma Revisited*; Cooper, *Cognitive Dissonance*.

presupposes underwrites our conception of impartial doxastic deliberation. However, there is reason to suppose that the Cartesian theory is more venerable myth than hard fact, and so we can be thankful that the Cartesian view is not the only available theory of belief formation. Spinoza had a competing view of belief formation, according to which contemplating a proposition's truth coincides with assenting to a proposition.[18] In lieu of the Cartesian view, I propose a version of a Spinozan theory of belief fixation, one in which tokening[19] a proposition is sufficient for believing that proposition. In 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).

What I am here calling the 'Spinozan Theory' 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. Because of our mental architecture, it is (nomologically) impossible for one to not immediately believe propositions that one tokens.

(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 sometimes refer to the Spinozan position as an 'asymmetrist' position, because it treats accepting and rejecting asymmetrically.[20]

(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

[18]Spinoza, *Ethics*.

[19]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 (importantly, as I understand them, these acts can either occur unconsciously or have an unconscious counterpart). 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 activated in one's LOT, one believes that sentence.

[20]The reader may see a certain affinity between the current claim and certain so-called dual process theories of reasoning. For example, Frankish, *Mind and Supermind*. However, the superficial resemblances are misleading. I argue that the unconscious states (putatively 'system 1' states) count as beliefs because they are the ones that act in law-like manners and cause behavior, whereas the conscious states (putatively 'system 2' states) do not seem to behave in any law-like fashion and so are unfit for psychofunctional theories. That said, no doubt this quick note gives short shrift to an important issue, just one that is beyond the scope of the current endeavor. See, however, Mandelbaum, 'Attitude, Inference, Association'.

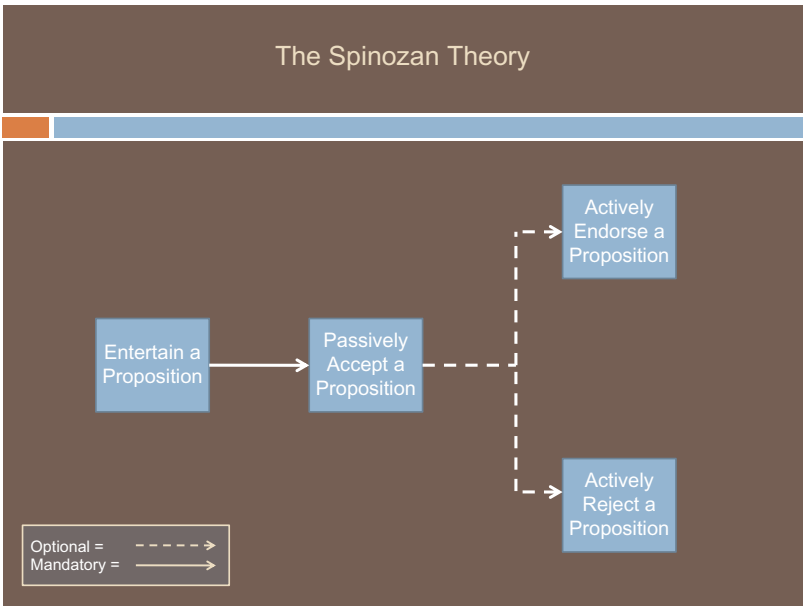

Figure 2. The Spinosan theory.

Note: The dotted lines represent optional links, and solid lines necessary links.

> accepted, but every proposition that is accepted is not necessarily endorsed.[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 the Spinozan theory dictates that accepting and rejecting are subserved by different mental processes, it is natural for such a theory to give some idea of what rejecting is. As opposed to analyzing rejection in terms of negation, I follow Price in analyzing negation in terms of rejection.[22]

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

The Spinozan view is consistent with the idea that, phylogenetically speaking, cognition grew out of perception. Since our perceptual faculties were by and large veridical, the cognitive faculties that first evolved just took the

[21]For current purposes, endorsing a proposition is something that happens at the person level. One consciously chooses what to endorse, whereas accepting need not 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 subpersonal phenomena).

[22]Price, 'Why "Not"'.

deliverances of perception at face value. The ability to reject information was a later evolutionary development. The Spinozan view sees this phylogenetic story mirrored in our current cognitive processing: just as the ability to accept information arose before the ability to reject information, so too do we automatically accept information before being able to reject it.[23]

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. For example, a Spinozan who denied property (4) could hold that accepting not-p does not entail rejecting p, though a Spinozan of my variety has to allow the entailment. However, no Spinozan can allow that one can reject p without also having first accepted p. Consequently, any Spinozan will predict that people (de facto) believe many contradictions.[24]

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 our current use, 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.

As per property (4), negating involves rejecting. Since negating involves rejecting, and since 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, the Spinozan theory predicts that negations can only be processed after a complete (affirmative) proposition has been formed. As a consequence, when processing negative clauses negations will be processed last.[25]

[23]Additionally, though developmental issues are beyond the scope of this essay, it seems that here ontogeny recapitulates phylogeny yet again, with the grasping of negation arising rather late in development. See, for example, Pea, 'Development of Negation'; Nordmeyer and Frank, 'Measuring the Comprehension of Negation'. Of course, one could accept the phylogenetic explanation (or the ontogenetic data) without the processing claim or vice versa, but it is the processing claim that I am most concerned with.

[24]Of course, this does not entail that people will *assert* contradictions. What one asserts is tied to what one endorses and on this picture endorsements are a species of judgment, not belief (for more on the relations between endorsing/denying and believing/rejecting, see the end of Section V.i) For the Spinozan judgments are a person-level phenomenon whereas beliefs are subpersonal.

[25]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

As a consequence of properties (2) and (4), 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 is quite natural for the Cartesian to be a symmetrist about negation as well as belief.

The big picture: on the Spinozan view, any propositional thought[26] 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 generate quite different predictions. For example, 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 occur under cognitive load. Additionally, because the Cartesian theory treats assenting and rejecting identically, it predicts that cognitive load will affect both processes identically.

By 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. Because the Spinozan theory posits that believing is reflexive, believing should occur even when one is under cognitive load. That is to say, 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 return to these predictions throughout the paper, but to foreshadow what is to come, the Cartesian theory faces a significant amount of problems. Experimental evidence from a wide range of sources casts serious doubt on the viability of theory.[27] Accordingly, what we need is an

after the clause sans negation (i.e. 'Jesse is a communist') is processed, rather than after the entire sentence sans negation ('John believe that Jesse is a communist') is processed. The theory is supposed to hold over both sentential ('I do not regret going to your party') and constituent negation ('I do regret not going to your party'). For constituent negation the VP is presumed to be processed before the negation, but the negation processed before the rest of the outlying sentence.

[26]'Propositional thought' just means 'thought that is truth-apt'.

[27]For some more anti-Cartesian data I do not have the space to cover, see Festinger and Maccoby, 'On Resistance'; Gilbert, Tafarodi, and Malone, 'You Can't Not Believe Everything You Read'; Gilbert, 'Inferential Correction'; Anderson, Lepper, and Ross,

alternative theory of belief fixation. I offer the Spinozan theory as a stark alternative to the Cartesian theory, one that is at least consistent with, if not outright supported by, the experimental evidence. However, my aim is not to defend the theory in full, but instead to articulate it as a viable theoretical alternative. The Spinozan theory will raise a host of questions and I have scant room to answer them all, but offering alternative cognitive architectures, architectures that we at least do not know to be false, is a crucial early step toward reaching a theory of the nature of belief. Toward that end, let us turn our attention to evidence that should make one wary of the Cartesian view.

III. The Case against the Cartesian

III.i. Memory asymmetries between truths and falsehoods

The quintessential 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 partake in a learning task while intermittently under cognitive load and are then tested about what they learned. (For example, they would learn about fictional criminal acts and then decide on an appropriate sentence, or they would learn about random people and then assess their mental states.) In one typical experiment, participants were asked to learn nonsense word meanings. They watched a computer screen on which sentences of the form 'An X is a Y' appeared, in which the 'X' was a nonsense word and the 'Y' was a word in English (e.g. 'A suffa is a cloud'[28]). Right after participants read a sentence, the screen flashed either the word 'true' or the word 'false', indicating whether the previous statement was accurate or not. Participants were also 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 was introduced in order to induce cognitive load. During the critical trials, participants read six true and six false claims. While reading four of these claims (two true, two false), the participants were interrupted by the tone. At the end of the trials the sentences were then turned into questions (e.g. 'Is a suffa a cloud?'), which the participants answered.

The Cartesian view predicts that the tone task should affect both true and false statements equally since, although contemplation has occurred, the participants have not 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

'Perseverance of Social Theories'; Kruger, 'Lake Wobegon Be Gone!'. All of these are covered in detail in Mandelbaum, 'Architecture of Belief'.

[28]From Gilbert, Krull, and Malone, 'Unbelieving the Unbelievable'.

belief-fixation 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.[29]

This type of asymmetry is evident across a range of experiments. A person 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. As the experiment above displays, accepting a proposition (i.e. remembering the proposition as true) comes much easier than rejecting a proposition (i.e. 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 100 by increments of 5 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, 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.

III.ii. Belief perseverance

Another telling set of experiments comes from the literature on belief perseverance in the face of experimental debriefing. For example, in a typical experiment, an experimenter asks participants to read a collection of suicide notes and to sort the real ones from the fakes. In a study by Ross et al.,[30] participants encountered 25 pairs of notes and were told that one note from each pair was a real note, the other a fake. After seeing each pair, participants would judge which note was real and which fake and were then given feedback on their performance. Participants were then (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

[29]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.
[30]Ross, Lepper, and Hubbard, 'Perseverance in Self-Perception'.

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. 24 out of 25 correct), they believed that they were better than average at the task, and if they received negative false feedback (e.g. 7 out of 25 correct), they believed they were worse than average at picking out real suicide notes from fake ones.

This experiment is not generally taken to illuminate anything about belief acquisition per se. It seems that the participants formed their beliefs in a reasonable 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 some beliefs are 'sticky'—hard to relinquish once formed. If so, then the debriefing effect would not tell us anything about belief acquisition per se, but rather belief perseverance.

But consider what happens if the people are briefed *before* they take part in the study and receive false feedback (we could 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 the formation of any beliefs about their ability if those beliefs are based on the bogus data. However, prebriefing the participants does not impact the participants' judgments about their ability. Wegner, Coulton, and Wenzloff replicated the Ross study except the participants were told *prior* to the task that the feedback would be dubious.[31] 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 persistence effects are anomalous on the Cartesian theory, casting doubt on a core commitment of the Cartesian view.

III.iii. Personality metrics

The next source of anti-Cartesian evidence comes from an unlikely area: personality psychology. Researchers in this area often present participants with a list of personality attributes and ask them to evaluate how much each attribute describes their personality. Consider a personality survey in which participants are given 20 statements and are asked to answer, for each statement, whether the statement applies to them or not. The participants answer 'yes' when the statement applies to them and 'no' when it does not. For 10 of the questions an answer of 'yes' corresponds with being an introvert and for the other 10 an answer of 'yes' corresponds with being an

[31]Wegner, Coulton, and Wenzloff, 'Transparency of Denial'.

extrovert. On such a scale, a 'perfect' introvert would be one who answered 'yes' to the 10 introversion questions and 'no' to the 10 extroversion questions, while a perfect extrovert would reverse their answers. When using such methods researchers have found that their data are sometimes compromised by 'yea-sayers'; that is, 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 20 questions, thus confounding the personality metric. The perfect 'nay-sayer' would reverse the pattern of the perfect yea-sayer.

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 to complete the survey, and both should be equally affected by cognitive load. Cartesian predictions are contraindicated by Knowles and Condon.[32] In their study, participants received a counterbalanced 100-item personality questionnaire and had their reaction times measured. Yea-sayers were operationalized as those who answered affirmatively on 53 or more of the items, and nay-sayers as those who answered affirmatively on 47 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. Yea-sayers take longer than appropriate responders, who take longer than nay-sayers. This response pattern is directly at odds with the first Cartesian prediction.

Cognitive load also affects yea-saying in a way that helps to disconfirm the Cartesian hypothesis. In a related study, participants were split into two groups, both of which were asked to answer 20 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 being asked to listen to the music and distinguish notes that came from the piano from those that came from other instruments. The non-loaded group heard the same sounds but was not asked to attend to them. The group under load was significantly more apt to answer affirmatively to the questionnaire, thus disconfirming the second Cartesian prediction. 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.

***

The evidence considered so far weighs against the Cartesian theory, and is not only compatible with the Spinozan theory but also helps support it. The memory asymmetry data are easily explicable if one assumes, as the Spinozan does, that when propositions are initially processed they are

[32]Knowles and Condon, 'Why People Say "Yes"'.

encoded as true and can only subsequently be marked as false by an effortful process that can be short-circuited by cognitive load. The prebriefing effect is also to be expected if one assumes a Spinozan architecture. When participants encounter the feedback, they automatically believe it, even though they know the feedback is false. Since they are engaged in a relatively fast-paced experiment, the participants lack the mental energy to override the false beliefs. Finally, the personality metric data are predicted by the Spinozan architecture. If negations are processed subsequent to affirmations, as the Spinozan view would have it, then we should expect nay-saying to take more energy, and thus more 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 first to 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 passively to 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. As we saw, these predictions were borne out.[33]

Despite the experimental support the Spinozan theory receives, it still faces difficulties. The theory as stated is far from complete. First, there may be alternative explanations of the experimental data besides the Spinonzan model. Second, the Spinzona model itself faces prima facie objections. To defend the Spinozan model further, both types of challenges must be met.

IV. How Else Can the Experimental Data Be Explained?

IV.i. Gullibility heuristic

Instead of using the presented evidence to support an architectural-processing story (namely, the Spinozan theory), one may be inclined to see it as support for an explanation that appeals to a pretty simple heuristic. Call such a heuristic 'the gullibility heuristic'. The gullibility heuristic is a (putative) rule that states that one should accept whatever one perceives as true.

[33]More grist for the Spinozan mill: nay-sayers tend to have high scores on the 'Need for Cognition' scale (a scale that ranks how much cognitive effort one is apt to engage in), and yea-sayers tend to have low scores, just as the Spinozan theory predicts. Cacioppo and Petty, 'Need for Cognition'. Those who acquiesce more often do so because they are not inclined to expend more mental energy, so they end up believing whatever they token and then never reconsider these beliefs. Those who nay-say do so because they want more mental exercise and thus are willing to expend more mental energy, making them more apt to reject their extant beliefs.

Heuristics are posited as cognitive shortcuts. Roughly, the idea behind heuristics is that the tougher the computational task, the more apt one is to use a heuristic (assuming one is available). If the problem one is dealing with is too computationally demanding (e.g. making a probability judgment), then one typically does not engage in the demanding processing and instead uses a rule of thumb (like trading in representative categories for probabilistic distributions).[34]

Figuring out what to believe is a very computationally demanding problem. It is 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 acquire (or update) and which to ignore based on one's current evidence, stock of beliefs, and recent actions.[35] Some have taken the problem to be so intractable that they see the scientific study of belief acquisition and updating (and central cognition in general) as a fruitless venture.[36]

The gullibility heuristic could be used to solve the Frame Problem. Perhaps what people do is initially believe everything most of the time, and then later toggle belief strengths in different ways.[37] In short, the problem of belief acquisition is exactly the type of problem that is ripe for a heuristic solution, so perhaps we should pursue that line of inquiry, and not look for an architectural solution. Furthermore, since many of the studies I have presented in support of the Spinozan theory depend on getting someone to believe some stimulus that is presented exogenously (e.g. the memory asymmetry studies), perhaps their results could be explained by merely positing the gullibility heuristic.[38]

Yet there are some strong reasons for doubting the gullibility heuristic hypothesis. For one thing, it cannot explain the belief perseverance effects. The participants in the belief perseverance studies have all the time they would like to form their own thoughts about their abilities. Furthermore, most heuristics can be 'turned off' or overcome in certain situations, especially when the participants are told that the heuristic they are using is inapplicable

[34]See Kahneman and Tversky, 'Framing of Decisions'.

[35]Dennett, *Brainchildren*.

[36]For example, Fodor, *Mind Doesn't Work That Way; Modularity of Mind*, 'Modules, Frames, Fridgeons'.

[37]Of course, this way of 'solving' the problem just pushes it one step back: now the problem will arise for updating beliefs as opposed to acquiring beliefs.

[38]The main modern proponent of Spinozan theories is Dan Gilbert, who proposed a forerunner to the view described here. If Spinoza deserves to be the namesake of this view, then Gilbert should at least be considered the modern intellectual progenitor of it. However, strictly speaking, Gilbert is just committed to the second and third properties of the Spinozan theory as I have described it here. Though he takes no specific stance on the architectural vs. heuristic question (and thus no stance on the question of the nomological *impossibility* of contemplating without believing, property [1]), it is most natural to read Gilbert as espousing this type of heuristic view. That said, in personal conversation he has told me that he is now sympathetic to the architectural story on offer here. See, for example, Gilbert, 'How Mental Systems Believe'.

to the situation at hand.[39] So, when participants are told that the feedback they are about to receive is false, they should override the putative gullibility heuristic and withhold from forming beliefs based on the dubious feedback (after all, they are explicitly being told that the situation they are about to encounter is one in which the heuristic does not apply). However, as we have seen, this is not how people behave. Thus, it seems like the belief perseverance effects are incompatible with the gullibility heuristic hypothesis.[40]

IV.ii. Informativeness objection

Some theorists have proposed that people do have the ability to entertain propositions without believing them, but only when they are entertaining propositions that are informative when they are false.[41] If this were the case, then all of the memory asymmetry evidence (Section III.i) used against the Cartesian view (and in favor of the Spinozan) would be undercut.

To support their hypothesis, Hasson et al. set up an experiment in which participants were given statements that were paired with faces. Upon seeing a statement, participants were told whether that particular statement was true or false. The experiment was designed so that 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

[39]See, for example: Nisbett and Ross, *Human Inference*; Chapman and Johnson, 'Incorporating the Irrelevant'.

[40]Even if we put aside the perseverance effects, there are still other insurmountable hurdles for the heuristic proposal. For example, people seem to believe everything they think, even when the ideas are *self-generated* and the participants are not under load. In a series of studies Epley and colleagues tweaked the traditional anchoring and adjustment paradigm (a subject to which I return in Section VI.ii) and showed that there are self-generated anchoring and adjustment effects. Since the gullibility heuristic says to believe what you *perceive*, it should only range over exogenously given stimuli, but the self-generated effects show that people believe endogenously created stimuli even when they know their creations are fatuous. Additionally, a heuristic explanation would predict that neurological damage should not cause a dissociation between acceptance and rejection, yet patients with damage to their prefrontal cortex do appear to show such a dissociation. In short, if the Spinozan architecture is accurate (and thus the gullibility heuristic off base), then what we should find are patients that have a deficit in rejecting information (since their rejection faculty is damaged) but we should never find patients who have a deficit in believing (assuming their abilities to comprehend are intact). That this appears to be what neuropsychology finds is rather heartening. Epley, 'Tale of Tuned Decks?'; Epley et al., 'Perspective Taking'; Epley and Gilovich, 'Anchoring-and-Adjustment Heuristic', 'Putting Adjustment Back'. See also Asp et al., 'Neuropsychological Test'; Asp and Tranel, 'False Tagging Theory'.

[41]Hasson, Simmons, and Todorov, 'Believe It or Not'.

participants were put under cognitive load.[42] 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: the cognitive load (interruption) had no effect on the recollection of true statements, but it did increase 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 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 seems 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.'[43] If their hypothesis were 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 were true, then people would not be able to entertain a proposition without believing that proposition *when the proposition is uninformative-when-false*. People could only entertain without believing when they are thinking about propositions that are informative-when-false. Prima facie, 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 to some extent consider P (such considerations need not be conscious). But what happens when you consider P? When you are 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 do not, but the informativeness hypothesis cannot seem to account for this fact. According to the informativeness hypothesis, if the proposition you are about to consider is uninformative-when-false you will believe it upon first hearing it, and if it is informative-when-false you will not, in which case you as perceiver of propositions have to be able to somehow see into the future to determine your disposition toward a proposition. But, of course, no one (*pace* Daryl Bem) wants to affirm a theory that entails parapsychological powers. There is

[42]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.
[43]Ibid., 568.

a way out of positing such powers and keeping the informativeness hypothesis, but it too is not the most promising route.

The reader may be thinking that Hasson et al. can just assume that we hold no relation when first encountering a proposition; that is, that we initially withhold assent like the Cartesian view supposes. But consider the following: it seems overwhelmingly plausible that before a person can determine how informative a proposition is, the person must first entertain the proposition (though, again, a person need not 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. The informativeness hypothesis dictates that people believe propositions after they have considered a proposition they have 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? If we have the ability to withhold assent, then why would we not use this ability in situations in which we are told a statement is false; why would we only use it when we are told a statement is false and that statement is informative-when false?

Showing that the explanation underwriting Hasson et al. is, to say the least, unclear still does not explain why they got the data they did. It would be nice if there were an explanation for what exactly caused these data to be generated. I now attempt to give you one that will happily be quite consistent with the Spinozan worldview.

I suspect that Hasson et al. ascertained their results because their study was 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 we would be because the latter is more shocking and vivid. When we encounter abnormal situations we are more apt to think longer and harder about the abnormal situation (in this example one might think: 'Who doesn't own a television?').[44] Finding out that someone does not drink tea for breakfast does not really get one's mental juices flowing, but finding out that someone does not own a television immediately raises some

[44]There is some evidence that deals with this line of thought. For example, the main thesis of Mandelbaum and Ripley 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. To repurpose an example from Uttich and Lombrozo, if you see me on a tuxedo at a fancy wedding, then you do not learn nearly as much about my mental states as if you see me in a tuxedo at the beach. See Brigard et al., 'Responsibility and the Brain Sciences'; Mandelbaum and Ripley, 'Explaining the Abstract/Concrete Paradoxes'; Uttich and Lombrozo, 'Norms Inform Mental State Ascriptions'.

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.[45] 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 in which one does not own a television for longer than they would consider the situation where one does not drink tea for breakfast (certainly the former would startle undergraduates, the study's participants, more than the latter). Accordingly, the subjects would perseverate on the thought 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 tell 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 does not drink tea for breakfast what comes to mind? There is no concrete mental image that occurs. However, when one considers someone who does not own a television, then many mental images pop up. In fact, one can see the difference in these statements as on par with the difference between the abstract and concrete innuendo effects. In studies of the perseverance of innuendos, we find that innuendos make a deeper impression when they are concrete rather than when they are abstract.[46] People can more easily ignore innuendos that are abstract (e.g. 'Audrey is not sour') than they can for innuendos that are concrete (e.g. '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.[47] 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 did not 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 is 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 quickly accessible negative counterpart.

Similarly, the informative statements in Hasson et al.'s studies can be easily envisioned when negated. When considering that this person does not 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

[45]Petty and Cacioppo, *Communication and Persuasion*.
[46]For example, Wegner, 'Innuendo and Damage to Reputation'.
[47]See Wason and Johnson-Laird, *Pscychology of Reasoning*.

old-timey radio, or a television that has been turned into a diorama). However, when I tell you that this person does not 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 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. help to support Spinozan view.

V. Objections to the Spinozan Model

V.i. Objection 1: introspection and intuition

Consider some fantastically odd proposition, like *dogs are made out of paper*. Are you not quite sure that you do not believe that dogs are made out of paper? And yet you entertained it. If your intuition is accurate, then the Spinozan theory must be wrong.

The intuition behind this type of argument is 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.[48] For example, 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.[49]

[48]For example, Bem, *Beliefs, Attitudes and Human Affairs*; Gopnik and Meltzoff, 'Minds, Bodies and Persons'; Lycan, 'Phenomenal Intentionalities', 'Tacit Beliefs'; Dretske, 'Naturalizing the Mind', 'Knowing What You Think'; Williamson, *Knowledge and its Limits*; Carruthers, 'Introspection', 'How We Know Our Own Minds'.

[49]Cooper, *Cognitive Dissonance*, 37. Lycan takes an even stronger line, asserting that beliefs are never conscious. Lycan writes, 'It is an interesting question whether we can ever introspect beliefs. On both phenomenological and theoretical grounds I doubt that too; what we introspect,

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 first recall your history with pintos. Perhaps you ordered them last week, so you infer you must like them or else you would not 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 need not engage in any elaborate inferential process: if I ask you whether you love your spouse, you generally know what the response is (or at least should be); 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 we love our spouse, or that genocide is abhorrent. 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 lead us to infer that all cases of belief are like our cases of strong 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, the last thirty years of psychology have shown how opaque our minds are to us. Philosophers have overplayed how much we can introspect because they often parochially focus on the contents of thought instead of on mental processes. Mental contents are sometimes available for report (of course, sometimes they are not, too); however, our thought processes are almost never available to report.[50] If they were available, then the cottage industry of priming studies would not be thriving. Not only are our mental processes unavailable for report, but we are even apt to misreport our emotional states.[51] It is reasonable to think that propositional attitudes are more like emotional states and mental processes than mental contents (and therefore just as difficult to report).[52]

One reason for thinking beliefs are unlike contents is because, although they have contents as proper parts, they are more than just contents: a belief is a content with a certain functional role, and to have a functional role is to *play a part* in our mental economy. Functional roles also should strike us as

in the way of cognitive items, are judgments, and we infer our knowledge of our beliefs from these.' Lycan, 'Tacit Beliefs', 64. Though I am sympathetic to Lycan's claim, the Spinozan hypothesis can rebuff the introspective objection with the weaker thesis in the text.

[50]See, for example, Nisbett and Wilson, 'Telling More Than We Can Know'.

[51]Dutton and Aron, 'Some Evidence for Heightened Sexual Attraction'.

[52]Even our metaphors for the attitudes (e.g. the 'belief *box*') show that beliefs are unlike contents —after all, the 'belief box' is supposed to be the 'place' you put certain contents.

reasonably similar to mental processes (both are operations on contents, not contents themselves) and as such should not be introspectable.

A slightly different way of arguing for the same point is to note that beliefs, like propositional attitudes in general, are *relations* to contents. Being able to introspect the content of beliefs does not imply the ability to introspect the beliefs themselves. To do that we would need to be able to introspect a certain relation to a content, and there is little reason to believe that we have introspective access to this relation. After all, these relations are to be spelled out in terms of the nomological connections of the belief states, and to figure out these connections we generally need to engage in empirical psychology.

Thus, in order to be able truly to introspect beliefs we would need to be able to introspect the functional roles and relations of beliefs. But it is difficult to see how one could introspect these dispositional and relational facts when even our empirical psychology has trouble ascertaining such facts.[53]

However, these philosophical arguments only go so far. The idea that beliefs can be merely introspected is deeply held, so it would be nice if there were some direct empirical evidence against the view that we can reliably introspect our own beliefs. Happily there is some evidence that speaks directly to the question of our access to our beliefs. Gweon, Young, and Saxe designed an experiment in which subjects formed beliefs about certain misleading pictures (e.g. a partially occluded picture would look like a fish, but once the occluder was removed it was revealed to be a picture of a snake) and experimenters recorded these judgments.[54] The subjects were then told what a different set of people believed about the pictures (e.g. that the partially occluded picture depicted a fish). Some of these beliefs turned out to be true, others false, and the subjects were aware of whether they had formed true or false beliefs. Fifty minutes later the subjects were asked to recall what they believed, and two very interesting results were found: first, subjects were *worse* at remembering whether they had true or false beliefs than they were at remembering whether others had true or false beliefs. Moreover, there is no evidence that the subjects felt as if they were misremembering—to them they were just telling the experimenters what they previously believed. Second, the same neural network that was activated when subjects were thinking about their own beliefs was activated when subjects were thinking about others beliefs. In particular the neural networks that were activated when thinking about one's own beliefs (or others) were the same that are activated in traditional 'theory of mind' experiments (the right and left temporo-parietal junction and the dorsomedial prefrontal cortex) Since it is clear that we do not have introspective access to others'

[53]For a similar argument put to a different use, see Goldman, *Simulating Minds*.
[54]Gweon, Young, and Saxe, 'Theory of Mind'.

beliefs, it is reasonable to suppose that these data are a strong indicator that we do not have introspective access to our own beliefs. But even if one wants to ignore the neural data, the behavioral data should speak for themselves.

In sum, I think it wise to remain unmoved by an objection that crucially relies on introspection. Rather, like Lycan, Dretske, and others, I suggest that 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, by inferring from what is reasonable to believe, or through some other investigative methods. Moreover, our lack of introspective access to our beliefs is central to solving multiple psychological puzzles: it is why there are so many implicit racists who make sincere avowals of their egalitarian beliefs; it is why people fall in love when traveling on other continents, mistaking fear for lust; it is why writing a counterattitudinal essay will sway what we report our beliefs to be. Many beliefs we think we do not harbor, we do; only we cannot figure that out merely through introspection—that is why we have clever psychologists.[55]

Returning to our original proposition, a ridiculous belief such as *dogs are made out of paper* is not a belief that is going to eventuate in much behavior, certainly not in the millisecond after contemplating it and before reporting you do not harbor the belief. Since this belief has such a low chance of causing any behavior, you could not come to find out that you harbor this belief even if you were excellent at reading your beliefs off of your behavior.[56] If you considered a more sensible though still outlandish proposition, such as *all dogs carry deadly viruses*, you would probably also claim not to believe it after consideration. But for all that, you would probably show subtle signs of harboring the belief. For instance, if after considering that proposition you were presented with dogs, you would probably start lightly sweating, the galvanic skin response being an effect of having considered the proposition. And note that, even if galvanic skin response is not the paradigm of intentional behavior, it is nonetheless behavior that needs to be explained, and the proffered hypothesis can do so.

When you consider a ridiculous proposition, you generally attempt to falsify the correspondingly acquired belief immediately. Assuming you are not under cognitive load, you can normally do this quickly. What then becomes available for introspection is your *judgment* that dogs are not made out of paper. From this you can rightly infer that you believe that

[55]None of this is meant to entail that we do not have a type of special first-person access to our beliefs; rather, it just implies that, if we do have such special access, it is not gained through a Lockean style of introspection.

[56]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 are not made out of paper, and we would expect stronger beliefs to win out (in most contexts) over weaker ones.

dogs are not made out of paper (after all, you cannot judge that X without believing that X).[57] Thus, in many of these cases people will both believe that dogs are made out of paper and believe that dogs are not made out of paper, but they will only think they have the latter belief because they have access to the judgment that accords with that belief.[58] 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 the idea that we do not have introspective access to our beliefs. Let me introduce Fred, a man in his early 30s who has always been pretty skinny and has imbibed a substantial amount of beer in his day. After turning 30, Fred became worried that he had a beer gut, although he did not actually have one. He would walk shirtless to the shower, and looking down would see a slight bulge in his belly, from which he inferred that he had a beer gut. After several weeks of this routine, Fred made a self-deprecating joke about his beer gut to his friends, who acted astonished at the suggestion. Fred then asked his friends if he had a beer gut and his friends said that he did not. Fred trusts his friends and believes that they are giving him a sincere response. His friends' adamant denials of the beer gut serve 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 Fred looks down at his stomach he *sees* a beer gut. Because he trusts his friends' words, Fred tries to discount these perceptions. For example, if Fred is asked if he has a beer gut, he asserts that he does not. One might thus reasonably suspect that Fred does not believe that he has a beer gut. Yet, if you want to predict the majority of Fred's behavior, your best bet is to believe that Fred believes that he has a beer gut. When Fred walks by a mirror, he is apt to turn sideways so see if he has a bulge; when Fred walks to the shower, he still looks down and gets a spike of anxiety; when Fred approaches the buffet table, he thinks twice about the fried chicken; when Fred sees a beer commercial, he winces; when Fred goes clothes shopping, he opts for baggy shirts instead of more form-fitting

[57]This follows on the tame assumption that conscious thoughts also involve tokenings in (for example) a language of thought. If so, then the inference from judgment to belief is secure even for the Spinozan.

[58]Of course, I now commit myself to the existence of contradictory beliefs, but I do so happily, for the uncovering of contradictory beliefs is legion in psychological inquiry. For some recent examples of contradictory beliefs see Cushman and Greene, 'Philosopher in the Theater'; Strickland et al., 'Syntactic Biases'; Ripley, 'Contradictions at the Borders'. That said, this commitment does not imply that people can have a single belief with self-contradictory content, only two separate beliefs whose contents are contradictory. For discussion, see Mandelbaum, 'Against Alief'.

ones. Yet Fred sincerely reports that he believes he does not have a beer gut. So, what is going on with Fred?[59]

One important datum in explaining this situation is realizing that Fred looks down and sees a beer gut much more frequently than he hears that he does not have a beer gut. Fred looks down and sees the beer gut every day, whereas Fred's friends' interventions happen quite infrequently. Although Fred discounts his beer-gut perceptions as optical illusions and he trusts his friends' reports that he is beer gut-less, he still acts as if he has a beer gut. The Spinozan theory proposes that Fred acts as if he believes he has a beer gut because he *believes* that he has a beer gut. Since Fred continually perceives that he has a beer gut, he is continually tokening the thought that he has a beer gut, and these tokenings are sufficient for believing that he has a beer gut. Moreover, the relative strengths of beliefs (between, say, believing one does have a beer gut versus believing one does not) are in part a function of how often each belief is activated.[60] Since Fred tokens the belief that he does have a beer gut more often than the belief that he does not, he believes that he does more strongly (and hence you see it in his behavior more clearly).[61] Nevertheless, he judges that he does not believe that he has a beer gut, because when he has discussed the issue in the past he has come to the sensible conclusion that he does not have one. However, since he cannot introspect his beliefs, he only reports the belief that seems most reasonable, which is his judgment that he does not have a beer gut.

Fred's case is by no means unique. The moral to be taken from such cases is that we need to make a distinction between belief reports and beliefs. What we can introspect are the former, and not the latter. Beliefs are unconscious propositional attitudes. By contrast, the Spinozan views belief *reports* as a species of judgments (a person-level phenomenon) that can be affected by all sorts of pragmatic factors. The beliefs that we report having are beliefs that

[59]Although I am a proponent of dissonance theory and think that its explanatory powers are often overlooked, dissonance will be of little use here because dissonance theory posits that people abhor inconsistencies. Dissonance theory would predict that Fred's behaviors would align with his assertions (or vice versa), in which case the person who keeps asserting he has no beer gut should start acting as if he did not have a beer gut.

[60]This is why the therapeutic advice of self-affirmation theory (saying what you want to believe over and over again) actually works. Steele, 'Psychology of Self-Affirmation'). It is strange to think that just saying over and over again 'I'm a good, smart, likeable person' actually makes a difference to one's beliefs about one's goodness, intelligence and likeability. That it does is a stark datum that is wholly explicable on the view I am offering. For a fairly comprehensive overview of this phenomenon, see Sherman and Cohen, 'Psychology of Self-Defense'.

[61]Additionally, every time Fred tokens 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. This situation raises the question of how negative thoughts can ever become stronger than positive thoughts. In short, the answer is partly based around the unsurprising observation that strength of belief is not just a function of tokenings.

on reflection we catalogue as normatively respectable. In essence, the beliefs that we report are the beliefs that we *endorse*, and we generally are wont to endorse only normatively justifiable propositions. Consequently, what we endorse is affected by a slew of heterogeneous factors, such as 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.[62] What we end up sincerely reporting as 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 endorse publicly.

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; at the unconscious level, beliefs—not judgments—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.

V.ii. Objection 2: but why are these states *beliefs*?

The last objection I consider is why we must conclude that the states I have been discussing are *beliefs* and not some other mental state. One might accept all the arguments I have given so far, and yet still not want to identify the states under discussion with belief. Perhaps one would rather identify them as aliefs or as some yet to be named mental state.[63] The motivation behind such a move is understandable: many of the epistemologically interesting properties of belief are missing from these states (these states are accepted arationally, you can hold contradictory ones, etc.). At this point, I could capitulate and accept that the view so far glossed is interesting enough without bestowing the term 'belief' onto these states. However, I do not think that calling these states beliefs is just to bestow some honorific upon them. Rather, I think that if we are to have an empirically adequate

[62]Evans, *Varieties of Reference*.
[63]Gendler, 'Alief and Belief', 'Alief in Action'; cf. Mandelbaum, 'Against Alief'.

theory of beliefs, then any of the conditions on what a belief must be will be met by the states under discussion.[64]

I take it that, whatever beliefs are, they must have at least the following properties: they must be semantically evaluable (have conditions of satisfaction), be able to be acquired by perception, interact with desires and other motivational states to cause behavior (so beliefs must have actual causal powers, in which case, one might want them to be mental particulars, *pace* Dennett), and perhaps most importantly, be able to serve as the premises in inferences (henceforth called 'inferential promiscuity'[65]). Although it would be nice if we could find an empirically adequate account of a mental state that met all of these conditions,[66] these last two conditions are non-negotiable. If we are to identify a given operationalized state as a belief, that state better be able to interact with motivational states to cause behavior and, in particular, it better be the type of state that can serve as the premise in inferences.[67] Beliefs can interact with other beliefs in order to generate new beliefs/knowledge. This last condition is not just a philosophical doctrine; it is this condition that has been used to separate full-blown beliefs from belief-like intramodular propositional states (such as our visual system's information that there is only one overhead light source).

Demonstrating that the states under discussion do have these properties would be an important step to showing that these states are belief-like. Equally important, it would be one of the only serious attempts I know of to identify beliefs with states that appear in empirical cognitive science.[68]

[64]Some of the issues here are difficult and involved; owing to space requirements I am only able to scratch the surface of some important debates. In particular I am not able to discuss issues about the metaphysics of belief, comparing analytic, teleological and psychofunctionalism. Suffice to say, that I think it is hard to square any analytic functionalism (or Dennett-style interpretationalism) with a desire to have an empirically adequate theory of belief.

[65]Stich, 'Beliefs and Subdoxastic States'.

[66]And it would be very nice if an account of such states could (a) explain how we can (more or less) believe anything we can assert; (b) analyze beliefs as two part relations (between a believer and something else, such as a proposition); (c) entail that beliefs can have the same content as other propositional attitudes (you can believe that there is a table in front of you, or desire it to be true); (d) explain the opacity of belief; (e) explain how beliefs give rise to Moore's paradox; and (f) mesh well with an empirical theory of mental processes. The first four conditions can be met if we assume that beliefs are relations to syntactically structured mental representations; this paper's argument is, in part, an attempt to meet this last condition. Fodor, 'Propositional Attitudes'.

[67]Of course, how this interaction works (and is implemented) is an empirical question. I see no reason why one must be committed to a picture where beliefs can only cause behavior through decision-theoretic means as opposed to heuristic (or other) means.

[68]Perhaps the most lauded recent attempt to do this is to be found in Schwitzgebel, 'Phenomenal Dispositional Account of Belief'. However, to my eyes Schwitzgebel's account runs afoul of two central desiderata: it cannot make sense of any of the acquisition data canvassed here and it is a deeply anti-realist account. It's Rylean dispositionalism is cleverly defended but by making beliefs dispositional it is unclear how beliefs are supposed to do any real causal work. Since it is unclear how dispositions can be proximal causes, it is unclear how beliefs can be understood

To start, notice that the states under discussion have clearly been caused by perception and are semantically evaluable; in all of the experimental designs the beliefs are acquired perceptually and generally given a truth value, so their semantic evaluability should not be in doubt. Oddly enough, even though this account divorces belief from assertion in many ways, the account can still give rise to Moore's paradox: one cannot felicitously assert 'p, but I don't believe p', for whatever proposition one entertains one believes and one has to entertain any state that one asserts (tokening by linguistic production mechanisms is just another way of entertaining).[69] But the really interesting work to be done is to show that these states cause behavior and are inferentially promiscuous. To do so, let us revisit the experimental paradigm introduced in Section III.i.

Lest one think that the asymmetry between remember truths and falsehoods holds just over 'mere memory' (whatever exactly that is supposed to be), perhaps one more example would help to show how this acquired information is used just like beliefs. In a study by Gilbert, Tafarodi, and Malone, participants were asked to watch a video-screen with two crawling lines of text on it, one on top of the other.[70] 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).

At the conclusion of the video, participants were asked to recommend a prison sentence for the offenses, ranging from 0 to 20 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 6 years when the false information was extenuating and 7 when it was exacerbating (not a significant difference), whereas their burdened counterparts recommended 5 years in jail in the extenuating condition and 11 years in jail in the exacerbating (a

actually to cause behavior. Since I take it that it is non-negotiable that beliefs must cause behavior, it is reasonable to suppose that Schwitzgebel's account has an important looming lacuna.

[69]That said, these states do allow a kind of inverse paradox to arise: one can assert 'not-p but I believe p'. Of course, if you sever the relation between belief and assertion, like I am inclined to do, the Moorean paradoxes are mere window dressing.

[70]Gilbert, Tafarodi, and Malone, 'You Can't Not Believe Everything You Read'.

statistically significant difference). Significant differences were also found across the board when looking at the defendant's likeability, ability to benefit from counseling and overall dangerousness. Thus, it seems that the burdened participants believed the lies they read more than the unburdened participants. Moreover, the falsehoods did not just get into the participants' 'belief box' and then were parroted out as responses; rather, the falsehoods became integrated with the participants' beliefs and affected a robust range of their responses. The propositions that the participants encountered while under load rippled 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 states which then informed their judgments concerning the duration of the sentence and the character's likeability. The attitudes the participants formed infiltrated and interacted with (presumably some subset of) their web of belief in order to produce the behavior the experiment detected.

The experiment just discussed is one of many (overlooked) studies that can help execute some real philosophical work. They serve as evidence that the states I have been discussing do indeed serve as premises in inferences and they show that the states are honest-to-god causally active in producing behavior. The inferential promiscuity is a very strong requirement—not any run-of-the-mill mental states can achieve it.[71]

In sum, it appears that the states we have covered do fit with our criteria for beliefs. Panning out a bit, at this point, we have seen that the experimental evidence counts against the Cartesian view, and is at the very least consistent with the Spinozan view. Moreover, some of the most glaring obstacles to the Spinozan theory have been removed. A final piece of support I offer for the Spinozan model comes from considerations about what other phenomena ballistic beliefs can help explain.

VI. Explanatory Power of Ballistic Beliefs

In this section, I briefly examine a subset of recalcitrant phenomena that ballistic beliefs could elucidate.[72] The unificatory and explanatory power of

[71]In particular, aliefs do not appear to be capable of inferential promiscuity. The content of an alief is not a truth apt proposition; in general, it is just a single mental representation. It is, to put it lightly, unclear how a single mental representation (such as DOG) can serve as a premise in an inference. There is much more to be said on the fecund topic of alief. For a discussion of what contents aliefs can take and aliefs' ability to pick out psychological kinds, see Mandelbaum, 'Against Alief'. See also Gendler, 'Alief and Belief', 'Alief in Action'.

[72]Some of the topics that the Spinozan view can help explain but which are not discussed here because of space limitations are the Fundamental Attribution Error (Jones, 'Rocky Road'), source monitoring errors (Sherman and Bessenoff, 'Stereotypes'); self-affirmation theory (Sherman and Cohen, 'Psychology of Self-Defense'); the 'fearing fictions' phenomenon (Walton, 'Fearing Fictions'); the efficacy of certain sorts of propaganda (Skinner, *Beyond Freedom and Dignity*); the 'abstract/concrete' paradoxes (Sinnott-Armstrong, 'Abstract +

the Spinozan theory gives us strong inductive reasons for believing the theory. One should read this section as an abductive argument. The Spinozan theory of belief formation is the only theory on offer that can give a unified explanation of what has previously appeared to be a slew of disconnected and problematic phenomena.

VI.i.'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 believe.[73] Part of the mystery of the confirmation bias is to explain why information that is consistent, as opposed to inconsistent, with a given standing belief comes to mind when we are assessing that standing belief. The bias is explicable on a basic dissonance theory.[74] A potted explanation goes something like this: if we believe in X and we find evidence that speaks against X, ascertaining such evidence will put us into a dissonant state. Since dissonant states feel bad,[75] they act as negative reinforcers, and through classical conditioning they reinforce us to not search for disconfirming evidence. This type of explanation articulates why the confirmation bias arises in the case of a previously held belief. However, the 'mere possibilities' version of the confirmation bias also arises in cases in which people are merely considering a proposition. For 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 respond that they are.[76] In these cases people search their memory for information that would confirm the question and then stop their search once they have reached such information. Dissonance theories have trouble explaining the mere possibilities formulation because they suppose that people are not yet invested in thoughts that they merely entertain. Thus, the mere possibilities formulation of the confirmation bias is a standing mystery. Note that this is not just a problem for dissonance theories; it is a problem for everyone since dissonance theories were the only real explanation of the confirmation bias. But the Spinozan theory can explain the mere possibilities version of the confirmation bias because the

Concrete ¼ Paradox'; Brigard, Mandelbaum, and Ripley, 'Responsibility and the Brain Sciences'); the ubiquity of implicit racism (Nosek et al., 'Pervasiveness and Correlates'); and the recovered memories phenomenon (Schacter, Norman, and Koutstaal, 'Recovered Memories Debate'). These topics are discussed in Mandelbaum, 'Architecture of Belief', Ch. 3.

[73]Lord, Ross, and Lepper, 'Biased Assimilation'; Klayman and Ha, 'Confirmation, Disconfirmation, and Information'.

[74]See, for example, Festinger, *Theory of Cognitive Dissonance*.

[75]Zanna and Cooper, 'Dissonance and the Pill'.

[76]Kunda et al., 'Directional Questions Direct Self-Conceptions'.

Spinozan theory entails that propositions that are merely entertained are thereby automatically believed. Since one believes merely contemplated hypotheses, the dissonance explanation can get a foothold and start doing its explanatory work.

VI.ii. Anchoring and adjustment

In the paradigmatic anchoring and adjustment paradigm,[77] experimenters ask participants to give numerical values in answer to some arbitrary questions, such as 'How old was Gandhi when he died?' and 'What is the freezing point of vodka?'[78] Before participants are allowed to answer the target question, the experimenter arbitrarily selects a number (e.g. by spinning a wheel, or by using a participant's social security number, or by a randomly chosen card), which serves as an 'anchor'. Participants are then asked whether the answer to the target question is higher or lower than the arbitrarily picked number. After answering this question, participants are allowed to answer the original question. The randomly generated anchors make a significant impact on the subjects' answers.[79] For example, people will guess that Gandhi died at 50 years old if they first have to decide whether he died before or after he was 9, and they will guess he died at 67 if they receive 140 as the anchor.[80]

Explanations of the anchoring and adjustment effect are scant at best. For example, the traditional 'explanation' of the effect is that people anchor onto a value and then adjust up or down from that value.[81] This explanation is just a restatement of the phenomenon. A more recent explanation of the effect is that it is produced by 'increased accessibility of anchor-consistent information'.[82] Although this seems like an explanation, this explanation itself is just an instance of a broader trend, the bias toward searching for confirmatory evidence; the confirmation bias. Hence, the confirmation bias (in particular, its mere possibilities version) is supposed to explain the anchoring and adjustment effect. But as we have just seen, an explanation of the mere possibilities version of the confirmation bias itself presupposes the Spinozan theory.[83]

[77]See, for example, Kahneman and Tversky, 'Judgment under Uncertainty'.

[78]Gandhi died at 78 years old; 80-proof vodka freezes at approximately −16.51°F.

[79]The participants given an anchoring number generally choose a number halfway closer to the anchors than the numbers chosen by participants who do not encounter an anchor. Jacowitz and Kahneman, 'Measures of Anchoring in Estimation Tasks', 1163.

[80]Strack and Mussweiler, 'Explaining the Enigmatic Anchoring Effect'.

[81]Kahneman and Tversky, 'Judgment under Uncertainty'.

[82]Epley and Gilovich, 'Putting Adjustment Back'; Mussweiler and Strack, 'Hypothesis-Consistent Testing', 'Numeric Judgment under Uncertainty'.

[83]One may be inclined to claim that anchor-consistent information becomes available through mere semantic priming. Maybe one's 'accumulator' (Gallistel and Gelman) is active when the number 140 comes up, and this primes other closer numbers. If this is so, then we would expect

As previously discussed, the confirmation bias is only supposed to be in play when one is 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 also needs to explain why merely contemplated hypotheses are believed. But the Spinozan theory can eliminate these mysteries.

Anchoring and adjustment effects arise because participants believe that the anchor they are given is actually the answer to the question they have been posed. Participants believe that the anchors are the correct answer merely because they entertained that possibility, and entertainment causes belief.[84] Note that, if one supposes a Cartesian theory, one cannot explain these robust effects without positing some very strange reasoning on the part of the believer.

VI.iii. Negation

A Spinozan theory that accepts property (4)—which states that to negate a thought is, in part, to reject it—makes many predictions regarding negation. Two in particular are germane to this short discussion: the prediction that negations are difficult to process and the prediction that negations are held back in the initial processing of a sentence. The former prediction is well known, so I keep my discussion of it short.[85]

VI.iii.i. Difficulty of negation

On the Spinozan theory rejection can occur only after acceptance. But it is not just the greater number of steps involved that makes rejection difficult; rather, it is that, since starting the rejection process is optional, one has to expend effort every time one rejects a proposition. The effort needed to reject a proposition is thus greater than the effort needed to accept a proposition. Since negations are a subset of rejections, applying a negation should also be an effortful, and thus difficult, task. This is a theoretical coup for the

---

that what participants do when they adjust is continually to slide along the number line until they reach a limit, one presumably dictated by the extent of the priming effect. However, Epley and Gilovich present data that speak against the sliding hypothesis and they propose instead that the adjustment phase is a series of jumps. Such jumps are inexplicable on the priming hypothesis. Gallistel and Gelman, 'Preverbal and Verbal Counting'; Epley and Gilovich, 'Putting Adjustment Back'.

[84]Here, as elsewhere, lies a tacit ceteris paribus clause. When participants are asked a question and then given the anchor, they must form a thought that turns the interrogative into a declarative. Fascinating evidence that participants do perform such a transformation can be found in Chapman and Johnson, 'Incorporating the Irrelevant'.

[85]For a more comprehensive overview of negation, see Mandebaum, 'Architecture of Belief'.

Spinozan because practically anywhere one looks, one can find data showing that negations are hard to process.

Adding negations to a sentence exponentially increases the difficulty in understanding the sentence with each additional negation. One does not need many data to see the point: it is easier to understand 'Jane didn't kick the ball' than it is to understand 'Jane didn't not kick the ball', which is much easier still than 'It is not the case that Jane didn't not kick the ball', and so on.

Negations also cause trouble when they are used as a search criterion. For example, people sort much faster and more accurately when they are asked to use a criterion that is positively formulated rather than negatively formulated.[86] Thus, people are much quicker at sorting when they are asked to sort out the spades and hearts from a pack of cards than when asked to sort out the non-clubs and non-diamonds.[87] We would expect both faster performance and fewer errors when using a criterion that involves less mental energy, and the Spinozan theory states that the processing of affirmations uses less energy than the processing of their negative counterparts.

VI.iii.ii. Psycholinguistic processing of negation

The second main prediction of the Spinozan theory regarding negation is more tendentious, though there is evidence that indicates 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 can 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 secondarily. This prediction is supported by Hasson and Glucksberg.[88] In their study, 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.

After reading the statements, the participants were shown a string of letters on a screen and were told to assess whether the letter string spelled an English word or not (i.e. they were given a lexical decision task). 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*

[86]Wason and Johnson-Laird, *Psychology of Reasoning*.
[87]Fodor, *Language of Thought*.
[88]Hasson and Glucksberg, 'Does Negation Entail Affirmation?'.

affirmative and negative sentences facilitated affirmative-related primes. However, the negative-related primes were *not* facilitated in the affirmative sentence conditions. 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.

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 does not take hold in processing until between 500 and 1,000 milliseconds 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.'[89] The Spinozan view (but not the Cartesian) predicts this startling psycholinguistic data.

****

In conclusion, the Spinozan hypothesis can help to explain quite disparate phenomena. The anchoring and adjustment effect and the confirmation bias are two of the most robust and mysterious psychological effects, and both can be elucidated using the Spinozan hypothesis, while the Cartesian model sheds no light on them. Additionally, the Spinozan theory can illuminate results about negation from many different branches of psychology—social psychology, cognitive psychology and psycholinguistics—while the Cartesian theory stays silent on the matter. Moreover, the Spinozan theory can explain one of the most well-supported findings in all of psychology: the fact that processing negation is hard. By contrast, the Cartesian model cannot even account for basic belief acquisition data. The Cartesian model continually misses the asymmetry between acceptance and rejection. For example, the Cartesian model makes all of the participants in the belief perseverance

[89]Ibid., 1027. For other work showing that affirmatives are processed immediately, see Blasko and Connine, 'Effects of Familiarity and Aptness'.

experiments look like exceptions even though their behavior is the rule. Surely, theories that can explain the relevant data are preferable to theories that claim that each datum is an exception. Thus, we should at least take the possibility of discarding the Cartesian hypothesis in favor of the Spinozan theory quite seriously. By doing so, we can explain and unify a number of mysterious phenomena in a theoretically respectable way. Even if there were no data against the Cartesian theory, this type of consilience should make one genuinely consider the merits of the Spinozan research program.

VII. Rationality Imperiled

Our conception of rationality is directly impacted by a Spinozan theory of belief acquisition. The Spinozan theory creates the following dilemma: either the ability impartially doxastically to deliberate is not a precondition on rationality, or people are 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 being able to justify our beliefs, and we have the expectation that these justifications are not just post hoc rationalizations. However, if the Spinozan theory is right, then we do not have the ability to deliberate about a proposition before believing it.

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 because of the confirmation bias (and the workings of dissonance reduction) we would also be unable to evaluate impartially the beliefs that we do hold. The confirmation bias (and dissonance effects) engender a biased deliberation strategy, one in which we tend to search for confirming information while ignoring disconfirming information. But this brings us to a depressing moral: at no point in our doxastic lives will we be able to consider propositions in a non-biased way. Yet it seems that our normative standards demand that a rational cognizer at least be able impartially to consider propositions at some point.[90] So, the first horn of the dilemma is quite unappealing.

The second horn is also unpalatable. For years research has been mounting that shows that people tend to be irrational in all sorts of domains; for example, we ignore base rates, we are 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

[90]If one wanted to deny that ought implies can (in epistemology), then our normative standards would not necessarily demand that we have the capacity to live up to them. This is not an unreasonable position, though it is not a way out of the dilemma either—it is tantamount to impaling oneself on the second horn.

of ourselves. The rational conception of ourselves is central to many theories of intentional ascription.[91] If we were to give up our conception of ourselves as rational creatures, then it is unclear what the paradigm of a rational creature would be.

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 Neurath's boat: not only are we are always reconstructing our boat at sea because we never have any fixed point from which to adjudicate our beliefs, but we also never have had any impartial perch initially impartially to form our beliefs. We are 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 toward confirming what we already believe, which is in part a function of whatever we happen to have entertained. We act as the epigraph from D'Alembert tells us to: we just start moving forward with whatever propositions we have encountered and our psychology makes it so that, often enough, the faith in those propositions comes too. And, of course, the propositions we happen to have encountered are often a hodge-podge. Sometimes a thought pops in one's head, not because of some reasonable inferential process, but instead because of one's dinner choice. And presumably we would not have wanted our epistemology held hostage to our gustation.

VIII. Conclusion

I have argued that there is a slew of evidence against the intuitive and ubiquitous Cartesian theory of belief fixation. As an alternative, I have offered a Spinozan theory of belief fixation. 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 a dearth of plausible theories of belief acquisition.

Acknowledgements

The first draft of this paper was written while I received funding from the American Council of Learned Societies, and revisions were made while I received institutional support through fellowships held at Oxford University, Yale University, and Harvard University. All of these institutions are hereby

[91]For example, Davidson, *Inquiries into Truth and Interpretation*; Dennett, *Intentional Stance*.

thanked for their generous support. The essay has also received a tremendous amount of feedback, for which I am grateful. Thanks are also owed to audiences at Georgia State University, Harvard University, the Massachusetts Institute of Technology, Oxford University, the University of Arizona, and the University of Waterloo. I received invaluable feedback from Molly Balikov, Tim Bayne, Ned Block, Ian Evans, Jerry Fodor, Uri Hasson, Bryce Huebner, Josh Knobe, Matt Kotzen, Bill Lycan, Shaun Nichols, Jesse Prinz, Mark Phelan, Dave Ripley, and Susanna Siegel. This essay is dedicated to the wonderful, late Fred Dretske, whose comments and support were utterly invaluable. Needless to say, the errors contained herein are all my own.

References

Anderson, C., M. Lepper, and L. Ross. 'Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information'. *Journal of Personality and Social Psychology* 39, no. 6 (1980): 1037–49.

Asp, E., and D. Tranel. 'False Tagging Theory: Toward a Unitary Account of Prefrontal Cortex Function'. In *Principles of Frontal Lobe Function*, ed. D. Stuss and R. Knight, 383–416. New York: Oxford University Press, 2012.

Asp, E., K. Manzel, B. Koestner, C. A. Cole, N. L. Denburg, and D. Tranel. 'A Neuropsychological Test of Belief and Doubt: Damage to Ventromedial Prefrontal Cortex Increases Credulity for Misleading Advertising'. *Frontiers in Neuroscience* 6 (2012): 1–9.

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

Blasko, D., and C. Connine. 'Effects of Familiarity and Aptness on Metaphor Processing'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19, no. 2 (1993): 295–308.

Bratman, M. 'Practical Reasoning and Acceptance in a Context'. *Mind* 101, no. 401 (1992): 1–16.

Brigard, F., E. Mandelbaum, and D. Ripley. 'Responsibility and the Brain Sciences'. *Ethical Theory and Moral Practice* 12, no. 5 (2009): 511–24.

Cacioppo, J. T., and R. E. Petty. 'The Need for Cognition'. *Journal of Personality and Social Psychology* 42, no. 1 (1982): 116–31.

Carruthers, P. 'How We Know Our Own Minds: The Relationship between Mindreading and Metacognition'. *Behavioral and Brain Sciences* 32 (2009): 121–82.

Carruthers, P. 'Introspection: Divided and Partly Eliminated'. *Philosophy and Phenomenological Research* 80 (2010): 76–111.

Chapman, G., and E. Johnson. 'Incorporating the Irrelevant: Anchors in Judgments of Belief and Value'. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. T. Gilovich, D. Kahneman, and A. Tversky, 120–38. New York: Cambridge University Press, 2002.

Cohen, L. *An Essay on Belief and Acceptance*. New York: Oxford University Press, 1992.

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

Cushman, F., and J. Greene. 'The Philosopher in the Theater'. In *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, ed. M. Mikulincer and P. R. Shaver, 33–50. Washington, DC: American Psychological Association, 2012.

Davidson, D. *Inquiries into Truth and Interpretation*. Oxford: Clarendon, 2001.

Dennett, D. *Brainchildren*. Cambridge, MA: MIT Press, 1998.

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

Descartes, R. *The Philosophical Writings of Descartes*. Cambridge: Cambridge University Press, 1988.
Dretske, F. 'Knowing What You Think vs. Knowing That You Think It'. In *The Externalist Challenge*, ed. Richard Schantz, 389–99. Berlin: Walter de Gruyter, 2004.
Dretske, F. 'Naturalizing the Mind'. Cambridge, MA: MIT Press, 1995.
Dutton, D., and A. Aron. 'Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety'. *Journal of Personality and Social Psychology* 30, no. 4 (1974): 510–7.
Epley, N. 'A Tale of Tuned Decks? Anchoring as Accessibility and Anchoring as Adjustment'. In *The Blackwell Handbook of Judgment and Decision Making*, ed. D. J. Koehler and N. Harvey, 240–57. Oxford: Blackwell, 2004.
Epley, N., and T. Gilovich. 'The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient'. *Psychological Science* 17, no. 4 (2006): 311–8.
Epley, N., and T. Gilovich. 'Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors'. *Psychological Science* 12, no. 5 (2001): 391–6.
Epley, N., B. Keysar, L. Van Boven, and T. Gilovich. 'Perspective Taking as Egocentric Anchoring and Adjustment'. *Journal of Personality and Social Psychology* 87, no. 3 (2004): 327–39.
Evans, G. *Varieties of Reference*. Oxford: Oxford University Press, 1982.
Festinger, L. *Theory of Cognitive Dissonance*. Palo Alto, CA: Stanford University Press, 1957.
Festinger, L., and N. Maccoby 'On Resistance to Persuasive Communication'. *Journal of Abnormal and Social Psychology* 68, no. 4 (1964): 359–66.
Fodor, J. *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press, 1998.
Fodor, J. *The Language of Thought*. New York: Thomas Y. Crowell, 1975.
Fodor, J. *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press, 2000.
Fodor, J. *Modularity of Mind*. Cambridge, MA: MIT Press, 1983.
Fodor, J. 'Modules, Frames, Fridgeons, Sleeping Dogs, and the Music of the Spheres'. In *Modularity in Knowledge Representation and Natural-Language Understanding*, ed. J. Garfield, 139–49. Cambridge, MA: MIT Press, 1987.
Fodor, J. 'Observation Reconsidered'. *Philosophy of Science* 51, no. 1 (1984): 23–43.
Fodor, J. 'Propositional Attitudes'. In *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, 177–203. Cambridge, MA: MIT Press, 1981.
Fodor, J. *Psychosemantics: The Problem of Meaning in Philosophy of Mind*. Cambridge, MA: MIT Press, 1987.
Ford, K., and Z. Pylyshyn, eds. *The Robot's Dilemma Revisited: The Frame Problem in Artificial Intelligence*. Norwood, NJ: Greenwood, 1996.
Frankish, K. *Mind and Supermind*. Cambridge: Cambridge University Press, 2004.
Gallistel, C., and R. Gelman. 'Preverbal and Verbal Counting and Computation'. *Cognition* 44, no. 1– 2 (1992): 43–74.
Gendler, T. 'Alief and Belief'. *Journal of Philosophy* 105, no. 10 (2008): 634–63.
Gendler, T. 'Alief in Action (and Reaction)'. *Mind and Language* 23, no. 5 (2008): 552–85.
Gilbert, D. 'How Mental Systems Believe'. *American Psychologist* 46, no. 2 (1991): 107–19.
Gilbert, D. 'Inferential Correction'. In *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. T. Gilovich, D. Kahneman, and A. Tversky, 167–84. New York: Cambridge University Press, 2002.
Gilbert, D., D. Krull, and M. Malone. 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information'. *Journal of Personality and Social Psychology* 59, no. 4 (1990): 601–13.
Gilbert, D., R. Tafarodi, and P. Malone. 'You Can't Not Believe Everything You Read'. *Journal of Personality and Social Psychology* 65, no. 2 (1993): 221–33.
Goldman, A. *Simulating Minds*. Oxford: Oxford University Press, 2006.

Gopnik, A., and A. Meltzoff. 'Minds, Bodies and Persons: Young Children's Understanding of the Self and Others as Reflected in Imitation and "Theory of Mind" Research'. In *Self-Awareness in Animals and Humans*, ed. Sue Taylor Parker, Robert W. Mitchell, and Maria L. Boccia, 166–86. New York: Cambridge University Press, 1994.

Gweon, H., L. Young, and R. Saxe. 'Theory of Mind for You, and for Me: Behavioral and Neural Similarities and Differences in Thinking about Beliefs of the Self and Other'. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 2492–7, 2011.

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

Hasson, U., J. Simmons, and A. Todorov. 'Believe It or Not: On the Possibility of Suspending Belief'. *Psychological Science* 16, no. 7 (2005): 566–71.

Huebner, B. 'Troubles with Stereotypes for Our Spinozan Psychology'. *Philosophy of Social Science* 39, no. 1 (2009): 63–92.

Jacowitz, K., and D. Kahneman. 'Measures of Anchoring in Estimation Tasks'. *Personality and Social Psychology Bulletin* 21, no. 11 (1995): 1161–6.

Jones, E. 'The Rocky Road from Acts to Dispositions'. *American Psychologist* 34, no. 2 (1979): 107–17.

Kahneman, D., and A. Tversky. 'The Framing of Decisions and the Psychology of Choice'. Science 211, no. 4481 (1981): 453–8.

Kahneman, D., and A. Tversky. 'Judgment under Uncertainty: Heuristics and Biases'. Science 185, no. 4157 (1974): 1124–31.

Klayman, J., and Y. Ha. 'Confirmation, Disconfirmation, and Information in Hypothesis Testing'. *Psychological Review* 94, no. 2 (1987): 211–28.

Knowles, E., and C. Condon. 'Why People Say "Yes": A Dual Process Theory of Acquiescence'. *Journal of Personality and Social Psychology* 77, no. 2 (1999): 379–86.

Kruger, J. 'Lake Wobegon Be Gone!: The "Below Average Effect" and the Egocentric Nature of Comparative Ability Judgments'. *Journal of Personality and Social Psychology* 77, no. 2 (1999): 1121–34.

Kunda, Z., G. Fong, R. Sanitoso, and E. Reber. 'Directional Questions Direct Self-Conceptions'. *Journal of Experimental Social Psychology* 29, no. 1 (1993): 63–86.

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

Lord, C., L. Ross, and M. Lepper. 'Biased Assimilation and Attitude Polarization: The Effect of Prior Theories on Subsequently Considered Evidence'. *Journal of Personality and Social Psychology* 37, no. 11 (1979): 2098–109.

Lycan, W. 'Phenomenal Intentionalities'. *American Philosophical Quarterly* 45, no. 3 (2008): 233–52.

Lycan, W. 'Tacit Beliefs'. In *Belief: Form, Content, and Function*, ed. R. Bogdan, 61–82. Oxford: Oxford University Press, 1986.

Mandelbaum, E. 'Against Alief'. *Philosophical Studies* 165, no. 1 (2013): 197–211.

Mandelbaum, E. 'The Architecture of Belief: An Essay on the Unbearable Automaticity of Believing'. PhD diss., University of North Carolina, Chapel Hill, 2010.

Mandelbaum, E. 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias' (manuscript).

Mandelbaum, E., and D. Ripley. 'Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The NBAR Hypothesis'. *Review of Philosophy and Psychology* 3, no. 3 (2012): 351–68.

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

Mussweiler, T., and F. Strack. 'Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model'. *Journal of Experimental Social Psychology* 35, no. 2 (1999): 136–64.

Mussweiler, T., and F. Strack. 'Numeric Judgment under Uncertainty: The Role of Knowledge in Anchoring'. *Journal of Experimental Social Psychology* 36, no. 5 (2000): 495–518.
Nisbett, R., and L. Ross. *Human Inference: Strategies and Shortcomings of Human Inference*. Englewood Cliffs, NJ: Prentice Hall, 1980.
Nisbett, R., and T. Wilson. 'Telling More Than We Can Know: Verbal Reports on Mental Processes'. *Psychological Review* 84, no. 3 (1977): 231–59.
Nordmeyer, A., and M. Frank 'Measuring the Comprehension of Negation in 2- to 4-Year-Old Children'. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, 3169–74, 2013.
Nosek, B., F. Smyth, H. Hansen, T. Devos, N. Lindner, K. Ratliff, C. Smith, K. Olson, D. Chugh, A. Greenwald, and M. Banaji. 'Pervasiveness and Correlates of Implicit Attitudes and Stereotypes'. *European Review of Social Psychology* 18 (2007): 36–88.
Pea, R. 'The Development of Negation in Early Child Language'. In *The Social Foundations of Language and Thought*, ed. D. Olson, 156–86. New York: Norton, 1980.
Petty, R., and J. Cacioppo. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer, 1986.
Pollock, J. *Contemporary Theories of Knowledge*. Totowa, NJ: Rowan and Littlefield, 1986.
Price, H. 'Why "Not"'. *Mind* 99, no. 394 (1990): 221–38.
Pylyshyn, Z. 'Computing in Cognitive Science'. In *Foundations of Cognitive Science*, ed. M. I. Posner, 49–92. Cambridge, MA: MIT Press, 1989.
Quine, W. V. *Word and Object*. Cambridge, MA: MIT Press, 1960.
Ripley, D. 'Contradictions at the Borders'. In *Vagueness in Communication*, ed. Rick Nouwen, Robert van Rooij, Hans-Christian Schmitz, and Uli Sauerland, 169–88. Heidelberg: Springer LNCS, 2011.
Ross, L., M. Lepper, and M. Hubbard 'Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm'. *Journal of Personality and Social Psychology* 32, no. 5 (1975): 880–92.
Samuels, R., S. Stich, and M. Bishop. 'Ending the Rationality Wars: How to Make Disputes about Human Rationality Disappear'. In *Common Sense, Reasoning and Rationality*, ed. R. Elio, 236–68. New York: Oxford University Press, 2002.
Schacter, D., K. Norman, and W. Koutstaal. 'The Recovered Memories Debate: A Cognitive Neuroscience Perspective'. In *Recovered Memories and False Memories: Debates in Psychology*, ed. C. Martin. New York: Oxford University Press, 1997.
Schwitzgebel, E. 'A Phenomenal Dispositional Account of Belief'. *Nous* 36 (2002): 249–75.
Sherman, D. K., and G. L. Cohen. 'The Psychology of Self-Defense: Self-Affirmation Theory'. In *Advances in Experimental Social Psychology* 38, ed. M. P. Zanna, 183–242. San Diego, CA: Academic Press, 2006.
Sherman, J., and G. Bessenoff. 'Stereotypes as Source-Monitoring Cues: On the Interaction between Episodic and Semantic Memory'. *Psychological Science* 10, no. 2 (1999): 106–10.
Sinnott-Armstrong, W. 'Abstract + Concrete = Paradox'. In *Experimental Philosophy*, ed. J. Knobe and S. Nichols, 209–30. New York: Oxford University Press, 2008.
Skinner, B. *Beyond Freedom and Dignity*. Indianapolis, IN: Hackett, 2002.
Spinoza, B. *Ethics*. 1677. Indianapolis, IN: Hackett, 1991.
Stalnaker, R. *Inquiry*. Cambridge, MA: MIT Press, 1984.
Steele, C. M. 'The Psychology of Self-Affirmation: Sustaining the Integrity of the Self'. In *Advances in Experimental Social Psychology* 21, ed. L. Berkowitz, 261–302. San Diego, CA: Academic Press, 1988.
Stich, S. 'Beliefs and Subdoxastic States'. *Philosophy of Science* 45 (1978): 499–518.
Strack, F., and T. Mussweiler. 'Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility'. *Journal of Personality and Social Psychology* 73, no. 3 (1997): 437–46.

Strickland, B., M. Fisher, E. Peyroux, and F. Keil. 'Syntactic Biases in Intentionality Judgments'. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 601–6, 2011.

Tuomela, R. 'Belief vs. Acceptance'. *Philosophical Explorations* 3, no. 2 (2000): 122–37.

Uttich, K., and T. Lombrozo. 'Norms Inform Mental State Ascriptions: A Rational Explanation for the Side-Effect Effect'. *Cognition* 116 (2010): 87–100.

van Fraassen, B. *The Scientific Image*. Oxford: Oxford University Press, 1980.

Velleman, D. 'On the Aim of Belief'. In *The Possibility of Practical Reason*, 242–82. Oxford: Oxford University Press, 2000.

Wallace, David Foster. *Everything and More: A Compact History of Infinity*. New York: W. W. Norton and Company, 2003.

Walton, K. 'Fearing Fictions'. *Journal of Philosophy* 75, no. 1 (1978): 5–27.

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

Wegner, D. 'Innuendo and Damage to Reputation'. *Advances in Consumer Research* 11 (1984): 694–6.

Wegner, D., G. Coulton, and R. Wenzloff. 'The Transparency of Denial: Briefing in the Debriefing Paradigm'. *Journal of Personality and Social Psychology* 49, no. 2 (1985): 338–46.

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

Zanna, M. P., and J. Cooper. 'Dissonance and the Pill: An Attribution Approach to Studying the Arousal Properties of Dissonance'. *Journal of Personality and Social Psychology* 29, no. 5 (1974): 703–9.