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

My dissertation examines how belief acquisition affects models of cognitive architecture and theories of rationality. In what follows, I detail the structure of the project and explain how its methods, arguments, and explanations make it a truly interdisciplinary dissertation and not merely a philosophical one. In its first chapter, I address the need to engage in interdisciplinary work to make substantial progress in cognitive science, and I explain how my work should be seen neither as purely philosophy nor as purely psychology. Although some may see the lack of disciplinary purity as a weakness, it is actually a boon: any work interested in understanding the human condition should have one foot firmly rooted in the sciences, the other in the humanities.

Throughout the dissertation, its interdisciplinary nature is manifest in both its method and content. The evidence adduced in support of my thesis is wide ranging and my arguments use premises that derive their warrant from social psychology, cognitive psychology, linguistics, neuropsychology, and philosophy. The subject matter and potential impact of my thesis also crosses interdisciplinary borders. Of particular relevance is my contention that talk of 'belief' in philosophy and the cognitive sciences has been ambiguous between the folk-psychological notion of belief, in which belief is seen as a consciously accessible, rationally acquired state, and the psychofunctional notion of belief, in which belief is characterized as an unconscious and arationally acquired state.[1] I argue that only the latter behaves in a law-like way, and as such is the proper subject for study in cognitive science. I then use the available data on the psychofunctional notion of belief (which stems from many different disciplinary origins) to motivate a novel model of cognitive architecture. I conclude by turning my attention to rationality. The cognitive architecture I propose opens up new avenues of research for situating the possible locus of rational control in the human mind, while denying that such control can happen at the traditionally supposed place: in our ability to contemplate propositions while withholding assent.

[1] Briefly stated, the psychofunctional theory implicitly characterizes beliefs by the empirical generalizations into which they enter.

OVERVIEW

Most models of cognitive architecture, that is models of the arrangement of mental faculties, focus on the question of how modular the mind is. However, in my dissertation I approach modeling cognitive architecture from a different route: instead of examining (e.g.) how many modules we can realistically posit, I focus on the motivation for central cognition itself through examining the way we actually acquire beliefs. The novelty of this approach is revealed through a comparison with classical approaches to cognitive architecture. In the traditional Fodorean architecture (henceforth *traditional modularity*), modules are at the periphery of the mind and more or less correspond to both our sense modalities and our language faculty. Traditional modularity posits these modules, which are informationally encapsulated, domain-specific processors, in addition to central cognition, a warehouse where our entire web of belief can exist. The postulation of central cognition is motivated by an understanding of belief fixation as a slow, rational, conservative process. Since traditional modularity theories assume that belief fixation is rational, they must propose a place where the totality of one's beliefs can be stored so that new information can be judged against one's stock of beliefs.

As traditional modularity has evolved, more modules have been suggested (e.g., for face perception, numerical cognition, theory of mind judgments, etc.), but the structural conception of central cognition has remained relatively stable. When central cognition has been attacked, it has generally been by those who are influenced by evolutionary considerations.[2] The evolutionary psychology movement meshes well with the idea of massive modularity, a model of cognitive architecture that rejects central cognition altogether. Evolutionary psychologists' eschewal of central cognition has generally been driven by what they perceive as a lack of compelling evolutionary stories explaining the genesis of central cognition.

My dissertation offers an independently motivated, irenic model of cognitive architecture, one that is unlike traditional modular models or massively modular models. My theory hypothesizes

[2] See, e.g., P. Carruthers, *The Architecture of the Mind* (Oxford: Clarendon, 2006); or S. Pinker, *How the Mind Works* (New York: Penguin, 1997).

that beliefs are stored in fragments that are created in an ad hoc fashion, fragments that are constantly being formed, merged, and morphed. Unlike traditional modular theories, this fragmented picture rejects a single central web of belief, and unlike massively modular models, it does not conceptualized these fragments as anything approaching modules: they are neither domain specific nor informationally encapsulated; rather the fragments are created and organized through the contingencies of one's experiential history, not by conceptually coherent domains.

Like those of the massive modularity theorists, my model eschews central cognition, but the motivation behind this choice is quite different. As opposed to most massively modular models, my theory does not rely on tendentious evolutionary considerations to advance its position. Instead, it challenges traditional modularity's assumption that belief fixation is actually slow, rational, and conservative. I argue that it is none of these things, thus undercutting the initial motivation for positing central cognition. I maintain that belief acquisition is an instantaneous, reflexive, and arational process. In essence my dissertation gives the first philosophical explication of Daniel Gilbert's suggestion that thinking is believing. Specifically, I maintain that merely entertaining that p suffices for believing that p.

The philosophical sharpening of Gilbert's provocative suggestion is necessary because there are many murky questions surrounding his formulation, such as: What are the properties that beliefs have in this picture? What is the relation between belief acquisition and rationality when beliefs are acquired in this ballistic fashion? What is the relation between a ballistic picture of belief formation and the cognitive architecture in which belief formation is situated? Does thinking suffice for, or cause, belief? In either case, is the relation maintained because of a heuristic or is it part of the basic architecture of the mind? My dissertation offers arguments in response to all of these questions. The position I ultimately advance can be summarized in the following way: merely (consciously or unconsciously) entertaining a (truth-apt) mental representation suffices for believing the proposition expressed by the mental representation, and this relation holds not because of any default heuristic

but because of the basic architecture of the mind.[3] A simpler, though less precise formulation: whatever one contemplates, one thereby believes.

DISSERTATION PRÉCIS

I begin my dissertation by describing and justifying the methodology I have adopted. Such a task, while sometimes tedious, is also necessary. After all, while I am a philosopher, my dissertation is an inherently interdisciplinary one that trades in both philosophical analyses of belief and rationality and also empirical models of belief fixation and belief storage. In the first chapter I explain that I am taking part in a long tradition of speculative psychology, a field that thrived until the 1940s. It isn't what we now think of as paradigmatic philosophy because it is concerned with empirical theory construction; it isn't paradigmatic psychology either because it doesn't present new experimental data. Instead, my methodology is to look at a slew of fascinating, recalcitrant, and seemingly disconnected data and attempt to explain them via an elegant and independently motivatable theory. The bulk of my dissertation undertakes this project, which I understand to be similar to the one in which the classical humanists partook and to be the central aim of truly interdisciplinary cognitive science. My inquiry into belief fixation is used as an inquiry into human nature—into the structure of the mind and why smart people often believe such stupid things.

In essence, chapter 1 addresses both why theoretical cognitive science counts as philosophy and why philosophers are qualified to engage in cognitive scientific theory construction. Moreover, it argues for why a person who is interested in the topics I am *must* do interdisciplinary cognitive science, even if he or she is bound to write a dissertation within a single discipline. It concludes by discussing why one might be interested in cognitive architecture even if he or she is simply a philosopher and not a cognitive scientist. As I see it, an end goal of such research is to find the locus

[3] I say 'truth-apt' because truth-apt mental representations are the only ones that are fit for belief: one cannot believe the open sentence …IS A DOG; that formula doesn't have the right structure to be believed, whereas FIDO IS A DOG does have the correct structure to be believed (small caps denote a structural description of a thought). Some further terminological notes: to entertain a mental representation is to occurrently access it or, in other words, think *with* it. A mental representation is accessed when it is used by a mental process. Neither the mental representation nor the process needs to be conscious for the mental representation to be accessed (thus I allow for unconscious entertaining, which I think is not an uncommon occurrence).

(or loci) of rational control, whether at the 'person-level' or 'sub-personally.' The end of chapter 1 discusses what types of control we could hope to discover, how to go about locating them, and what it would mean to find such loci of control.

Following chapter 1's discussion of methodology, the real model-building begins in chapter 2. I start by sketching the Cartesian theory as it appears in Descartes's model of the mind and tracing it through contemporary accounts of the mind. Descartes purports to withhold his assent from all beliefs that are not 'clear and distinct,' contemplating propositions without actually believing them.[4] Descartes's attempt presupposes a serial model of the mind, one where propositions can be merely entertained and judgment can be suspended. Contemporary cognitive science has elaborated this model by distinguishing modular systems from central systems.[5] Modular systems automatically, unconsciously, and instantaneously provide information about how the world appears, and central cognition reflects on this information and either accepts or rejects it (or suspends judgment). The canonical modular model of the mind thus presupposes the possibility that one can process information without automatically believing it.

In both the earlier Cartesian and more contemporary modular models, belief acquisition is analyzed as a slow, conservative, controlled, serial process, one where propositions can be considered before acceptance or rejection occurs (both attitudes being underwritten by the same mental process). Because it assumes that propositions can be considered and then assented to or rejected, the Cartesian system allows for beliefs to be formed based on rational evidence; thus belief acquisition can be understood as a rational process. In contrast, I, in the spirit of Baruch Spinoza and Daniel Gilbert, propose that belief fixation is an arational, reflexive process.[6] The basic propositions defended in chapter 2 are that a) it's impossible to initially withhold assent from an entertained

[4] R. Descartes, *The Philosophical Writings of Descartes* (Cambridge: Cambridge University Press, 1988).
[5] J. Fodor, *The Modularity of Mind* (Cambridge, Mass.: MIT Press, 1983).
[6] B. Spinoza, *The Ethics and Selected Letters*, ed. S. Feldman and trans. S. Shirley (Indianapolis, Ind.: Hackett, 1982); D. Gilbert, 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.

proposition; b) acceptance is underwritten by a different mental mechanism than rejection;[7] c) forming a belief occurs passively, while rejecting a formed belief is an active and effortful endeavor; d) the acceptance of a proposition necessarily occurs before any rejection can take place; and e) the properly sanitized notion of belief employed by the laws of cognitive science characterizes belief as an unconscious, unintrospectable attitude, one that fulfills conditions a through d.

The bulk of chapter 2 presents evidence that the Cartesian view is false. The evidence adduced in this argument includes, inter alia, memory asymmetries between remembering truths and falsehoods, belief perseverance in the face of experimental debriefing, cognitive load's effects on yea-saying and nay-saying, and the efficacy of counter-attitudinal communications while under load. In sum, chapter 2 compiles a suite of phenomena that contradicts the Cartesian theory yet is naturally explicable on the 'thinking is believing' model.

After motivating the model described in chapter 2, chapter 3 then shows what one can do with it. The chapter as a whole serves as an elongated abductive argument, one that proceeds by exploring the explanatory fruits of the thesis. The simple proposition that thinking is believing has the power to elucidate hitherto disparate and mysterious effects. These effects include: the fundamental attribution error;[8] the 'mere possibilities' formulation of the confirmation bias;[9] the anchoring and adjustment effect;[10] the relationship between one's 'need-for-cognition' and one's disposition to acquiesce; the mechanisms whereby false recovered memories feel veridical;[11] the

[7] I also argue that negation involves the propositional attitude of rejection, a lemma that appears intermittently throughout the dissertation.
[8] In the fundamental attribution error, situational constraints are incorrectly interpreted as character traits. E. Jones, "The Rocky Road from Acts to Dispositions," *American Psychologist* 34, no. 2 (1979): 107–17.
[9] In such situations, people are given a hypothesis in which they have no prior investment, and yet they act as though they believe the hypothesis, searching for confirming, not disconfirming, information. J. Klayman and Y. Ha, "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review* 94, no. 2 (1987): 211–28.
[10] This effect is demonstrated when people are merely presented with numbers they know to be randomly derived before making a numerical estimation and yet the arbitrary numbers significantly affect the people's subsequent numerical judgment. D. Kahneman and A. Tversky, "Judgment under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (1974): 1124–31.
[11] D. Schacter, 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, 63–99 (New York: Oxford University Press, 1997).

efficacy of self-affirmation;[12] problems of stereotype fulfillment;[13] why negations are difficult to process and why they are processed last in language comprehension;[14] ironic innuendo effects;[15] how it's possible for people to 'fear fictions' and willfully suspend disbelief (when, e.g., watching a horror movie);[16] and implicit racist phenomena.[17]

My fourth chapter further clarifies the thesis and responds to objections to it. I begin by separating my thesis from the previous theories that have influenced my thinking. In particular, I argue that Gilbert, who holds a similar view, is actually committed to a heuristic model, one which claims that people follow a tacit rule: when under load, they should err to believe whatever they hear or see. Behind this claim is the thought that, in general, our perceptual faculties are veridical, so when we are under load it's a good rule of thumb to believe the deliverances of our perceptual faculties. I argue that this heuristic view cannot be sustained, since the effects of belief perseverance hold even when the propositions are self-generated and not the result of a perceptual process.[18] I then conclude that the effects must be part of the architecture of the mind.[19] Next, I specify the architecture,

---

[12] E.g., why repeatedly telling yourself you are competent makes you believe that you are. C. M. Steele, "The Psychology of Self-Affirmation: Sustaining the Integrity of the Self," in *Advances in Experimental Social Psychology*, vol. 21, ed. L. Berkowitz, 261–302 (San Diego: Academic Press, 1988).

[13] E.g., why women do worse at math exams when reminded of their gender before taking the exam. D. K. Sherman and G. L. Cohen, "The Psychology of Self-Defense: Self-Affirmation Theory," in *Advances in Experimental Social Psychology*, vol. 38, ed. M. P. Zanna, 183–242 (San Diego: Academic Press, 2006).

[14] For an example of the difficulty of processing negations, consider why 'May isn't June' is easier to understand than 'It's not the case that May isn't not June.' I argue that negation is difficult to process because it involves rejection, which is an effortful cognitive process. Because one can only reject a whole proposition, negations must be processed last in sentence comprehension, once the entire proposition has been formed. For evidence that negation is processed last in sentence comprehension, see U. Hasson and S. Glucksberg, "Does Negation Entail Affirmation? The Case of Negated Metaphors," *Journal of Pragmatics* 38 (2006): 1015–32.

[15] E.g., why saying 'Barack Obama is not a Muslim' leads people to believe he is a Muslim. D. Wegner, G. Coulton, and R. Wenzlaff, "The Transparency of Denial," *Journal of Personality and Social Psychology* 49, no. 2 (1985): 338–46.

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

[17] In such phenomena, people who explicitly aver egalitarian principles still appear to harbor anti-egalitarian sentiments when tested implicitly. See, e.g., D. Carney, N. Krieger, and M. Banaji, "Implicit Measures Reveal Evidence of Personal Discrimination," *Self and Identity* 9, no. 2 (2010): 162–76.

[18] See, e.g., the self-generated version of the anchoring and adjustment heuristic: N. Epley 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): 391–96.

[19] When I say 'part of the architecture' what I mean is that the procedure is built into the system; the system needn't work by checking against an unconscious rule in order to properly function. (For example, concluding Q from IF P THEN Q and P is something that our minds are set up to do [in certain circumstances], because the inference is primitively compelling. For arguments supporting this interpretation of modus ponens, see Fodor,

discussing not only how beliefs are acquired but also how they are stored in central cognition. My framework for belief storage is a 'fragmented network'; I argue that there is no single web of belief but instead ad hoc fragments that are created on the fly. Each fragment itself aims for consistency, but consistency is not achieved between fragments. My model thus explains a fact that is indisputable yet nearly impossible for contemporary epistemology to explain: people persistently hold contradictory beliefs.

The rest of the chapter deals with a variety of putative counterexamples to my view. In addressing these objections, I employ a wide range of evidence from across the cognitive sciences, from cognitive neuropsychology (e.g., data from Capgras patients) to philosophical considerations and analysis (e.g., I argue that the states I discuss must be beliefs because they have propositional structure, serve as the premises of inferences, and guide behavior, all properties that are central to traditional analyses of belief).

One of the numerous objections I address contends that my theory runs counter to people's intuitions: people feel that they are able to first consider and then reject propositions. I counter this objection by arguing that it presupposes that people have introspective access to their beliefs. Although such an assumption appears benign, it is unsustainable. I argue that beliefs are relations to mental representations, relations to be spelled out by the functional role of belief, which includes beliefs' roles in our mental economy (e.g., their relation to inference, action, perception, desire, etc.). But these relations are not themselves introspectable—if they were then we wouldn't need to engage in empirical psychology to discern them! Thus, regardless of one's preference on the question of belief acquisition, one should conclude that beliefs are not introspectable. I then note that this conclusion shouldn't be surprising, because most mental processes and propositional attitudes appear to exist beyond the reach of introspection. There is evidence that our intuitions of introspectability

---

The Modularity of Mind; L. Carroll, "What the Tortoise Said to Achilles," *Mind*, no 4 [1895]: 278–80; and C. Peacocke, *A Study of Concepts* [Cambridge, Mass.: MIT Press, 1995].)

are misguided when it comes to emotional states and mental processes,[20] so the fact that we are incorrect about our access to beliefs should come as no great surprise. People are misled to think that they can introspect their beliefs by that fact that they do have access to belief contents, yet a content is, of course, not a belief. I end by noting that we can know what content we are thinking about without knowing how exactly we are thinking about that content.

At this point in the dissertation, I take stock of the theoretical implications of the discussion. The big picture is that on the view defended in my dissertation 'belief' is seen as ambiguous between the folk-psychological notion, the notion typically employed in normative epistemology, and the cognitive scientific notion, in which belief is an unconscious, basic propositional attitude that compels behavior. One of the main conclusions of the second chapter is that it is the latter conception, belief as unconscious attitude, that actually enters into law-like relations, while the former conception, belief as it is used in folk psychology, does not enter into such nomic generalizations. Thus, according to my view, belief is not eliminated from our cognitive science, but it is transformed, from an introspectable, rational propositional attitude to an unconscious, arationally acquired propositional attitude. In sum, I conclude that philosophers and many cognitive scientists have been playing fast and loose with belief: they've ignored the empirical evidence for what beliefs actually look like. By taking into account the empirical evidence we can thus usher an empirically respectable notion of belief into our cognitive science. However, the resulting notion of belief looks a bit different than belief as it is seen in folk psychology. In particular, its acquisition and storage conditions greatly differ from the folk psychological notion of belief. What we end up with is no less than a proposal for a reconceptualization of the notion of belief, as well as the novel picture of central cognition that results from such a reconceptualization.

In my final chapter I use my model to reexamine both the notion of human rationality and epistemic norms. I argue that the model poses a new problem for our conception of rationality, one

[20] R. Nisbett and L. Ross, *Human Inference: Strategies and Shortcomings of Human Inference* (Englewood Cliffs, N.J.: Prentice Hall, 1980); and D. Dutton and A. Aron, "Some Evidence for Heightened Sexual Attraction under Conditions of High Anxiety," *Journal of Personality and Social Psychology* 30, no. 4 (1974): 510–17.

that differs greatly from those stemming from the 'rationality wars.'[21] My theory creates the following dilemma: either the ability to impartially doxastically deliberate is not a precondition of 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 academics we pride ourselves on being able to justify our beliefs, and we have the expectation that these justifications aren't post-hoc rationalizations. We expect to have some rational control over what we believe, even if this control is not 'top-down.' However, if my theory is right, then we don't have the ability to deliberate about a proposition before believing it.

That's just the start of the trouble for impartial deliberation. If my theory is correct, not only would we be unable to withhold assent from propositions, but we would also be unable to impartially consider the beliefs that we do hold. Because of the confirmation bias, we would have only a partial deliberation strategy, one in which we tend to search for confirming information while ignoring disconfirming information. Thus, at no point in our doxastic lives would we be able to consider propositions in a non-biased way. This is troubling since our normative standards demand that a rational cognizer be able to impartially consider propositions. Of course, the other horn of the dilemma, accepting that people are necessarily irrational, is just as unappealing. After all, rationality is about us—we are the paradigmatically rational creatures. If we are necessarily irrational, then our concept of rationality has no referent whatsoever. These considerations compel me to conclude that if we are to find the locus of rational control in human cognition, we cannot look to belief acquisition for help; instead we may focus our investigations on our methods of belief storage and the way in which beliefs change their strength.

[21] R. Samuels, 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).