The Self in the Age of Cognitive Science: Decoupling the Self from the Personal Level*

Robert D. Rupert
University of Colorado, Boulder

1. Introduction

This paper explores the nature of the self, in particular, what the self appears to be in the age of cognitive science. It does so through the lens of a commonly made distinction between the personal and the subpersonal levels of reality; for, given the supposed nature of the personal level, the personal level would seem the most likely home for the human self. Much of the discussion to follow takes a negative tone. I argue that, when one adopts the perspective of contemporary cognitive science, the personal level fades from view, and thus that we shall not find the self there; for there appears to be no such level. This negative tack produces positive results, however, by yielding a more accurate picture of the structure of cognitive science. Moreover, attention to the structure of working cognitive science promises to deliver insights into the nature of a human self as it is likely to appear from the perspective of contemporary cognitive science.

The paper proceeds as follows. After articulating the personal-subpersonal distinction, I lay out what I'll call the 'Received View' of the relation between the personal level and cognitive science. According to this view, the personal level is populated by process-

* This paper is a revised version of the transcript of a talk given in March, 2018, at the College at Brockport, SUNY. Many thanks to Joseph Long and the College's Center for Philosophic Exchange (CPE) for the invitation to speak and to the members of the CPE audience for their stimulating feedback. The paper builds on work that has long been underway and has been presented at the Universities of Edinburgh, Stirling, Sheffield, Macerata, Paris-Sorbonne, and Colorado; Ruhr University, Heinrich Heine University, Central European University, the Hungarian Academy of Sciences, a meeting of the Society for the Metaphysics of Science, and the University of Salzburg's Winter School in Cognitive Neuroscience. Enormous thanks to organizers and audiences at all of these venues. The list of philosophers and cognitive scientists discussions with whom have improved the final product (such as it is) is unmanageably long, and the generation of it, impractical. Nevertheless, I would like to thank explicitly Zoe Drayson and Bryce Huebner for exchanges that got me started on this project and Andy Clark, Mike Wheeler, and Alistair Isaac for their feedback along the way.

es, states, and properties about which much is known independently of science; for its part, cognitive science investigates the mechanistic, or more broadly physical, basis of such processes, states, and properties, which basis appears at the subpersonal level. The Received View places the self in a realm distinct from the processes and states investigated by cognitive science and assigns philosophers a distinctive role, as explorers of that distinct realm, by dint of their facility with *a priori* reasoning, conceptual analysis, and introspection. Cognitive science retains a modest role, determined for it from outside of cognitive science itself: studying how the physical organism implements, grounds, or enables the appearance of the personal-level facts, themselves known nonscientifically.

I argue that this picture radically misrepresents the structure of cognitive science. As an experimental science, its job is, first and foremost, to model publicly observable, replicable data. (And to the extent that cognitive science is a field science, its job is to model data collected according to the standards operative in the collection of data pertaining to, e.g., historical geological processes or to the spread of diseases through populations.) Examination of the working science itself seems to reveal no evidence of a personal level, and certainly not one the accounting for which provides the primary goal of, and a significant constraint on, cognitive science. Nowhere in the literature have I found a case in which working cognitive scientists rejected an otherwise superior model of extant data in favor of an otherwise inferior model of that same data because the latter model, but not the former, fit with *a prioristic* or otherwise armchair-generated claims about a supposed personal level. To the extent that cognitive science vindicates (or will, in the future, vindicate) commitments of *a prioristic* or common sense based reflection, it does so (or will do so) on its own terms, only where the entities, states, or distinctions in question appear in a significant range of cognitive science's best models of the relevant data, where, by 'best', I mean those that rise to the top given normal scientific standards having nothing to do with facts about a supposed personal level. I support this alternative view with two case studies.

It follows, then, that the relationship between cognitive science and the self is not what the Received View would suggest. It is not the case that cognitive science yields an account of the self by succeeding in its task of vindicating *a priori* (or introspection-based, or common sense based) commitments made independently of the science and meant to set the standard for success in cognitive science's investigation of the self or person. Rather, if cognitive science yields an account of the self at all, it does so (or will do so) by generating models that contain self-like constructs in them, that is, constructs that have properties or play roles sufficiently like the pretheoretic concept of the self as to warrant the application of that term. In closing, I speculate about the likely nature of such constructs.

2. The Personal and Subpersonal Levels

What, then, is the personal level supposed to be? The literature contains a variety of characterizations, and I shall not attempt to formulate a single, definitive picture or a set of necessary and sufficient conditions. Rather, in what follows, I content myself with a listing of properties typically associated with the personal level and some scattered remarks on potential connections between various of these properties.

Some authors identify the personal level with the realm of entities picked out by literal use of personal pronouns. They tend to write such things as, "My visual cortex doesn't see the tree. *I* see the tree," to indicate their conviction that perceiving is a state of a whole person (see, for example, McDowell 1994). (And sometimes such remarks seem to amount to little more than what I have in the past labeled the 'argument from the italicized pronoun'.) Importantly, though, such authors do not mean to claim something only about parts and wholes of organisms—that, for instance, because the whole organism has a property not shared with some of its parts, a distinct level of reality thereby exists. After all, a human's arm has a length but no height (and her body has height but no length). The fact that the whole body has a property not had by one of its parts (and vice versa) hardly entails the existence of a new level of reality. What might be added, then, to create an additional level of reality?

More often than not, consciousness serves as the additional ingredient. Humans are consciously aware of their thoughts, feelings, and sensory experiences, and such states seem central to who we are, central to our conception of ourselves as persons. Moreover, consciousness is often thought to carry with it special epistemeic status, providing specially secure knowledge available only from the first-person perspective. The individual is immediately acquainted with his or her own conscious experiences, in a way that no one else can be. The sheer uniqueness of this phenomenon suggests that conscious states have a distinctive ontological status.

Personal-level states are also thought to have a distinctive sort of content (McDowell 1994). The special nature of this content might consist in *(a)* its being propositional, *(b)* its being consciously accessible, or *(c)* its being built into conscious states themselves. Whereas subpersonal states function as mere detectors or causal mediators—like dummy lights on an automobile's dashboard—personal-level states such as beliefs can genuinely represent—and misrepresent—the world as being a certain way and can be offered by the subject as reasons in support of her beliefs (Brandom 1998).

Humans appear to act for reasons—that is, on the basis of what they think, believe, know, and desire—and it has been widely suggested that this form of reasons-based agency is distinctive of the personal level. Everyday explanatory practice—the workings of so-called folk psychology—reflects this fact. Folk psychology represents subjects' behavior as a product of their beliefs and desires, and the intentions formed on the basis of them. Why did Sarah go to the kitchen? Because she wanted a sandwich, and she believed that sandwich-makings awaited in the kitchen.

This way of understanding the personal level (and the persons who populate it) dovetails with another important strand of thinking about the personal level: that it is the realm in which one finds the kinds of states appealed to by the folk—beliefs, desires, and intentions, in particular.[1]

Rational coherence is often associated with the personal level. Although the human subconscious may be fragmented—filled with conflicting urges, fears, and inclinations—the entities populating the personal level are thought to maintain a certain consistency and to operate in accordance with certain logical rules (Colombo 2013, 549). At the personal level, one's psyche hangs together, it is thought, with no contradictory beliefs and no sets of intransitive preferences; if errors of rationality do appear at the personal level,

[1] As Rey (2001) sees things, once one recognizes the full range of states invoked by everyday psychological explanation—hopes, dreams, fears, states of surprise, convictions, annoyances, joy, grief, anger, excitement—the placement of folk psychological states at the personal level, as a domain distinct from the level investigated by workaday cognitive science and distinguished particularly by its being the home of belief-desire explanation—loses its plausibility. Here Rey builds on the insights of Hursthouse (1991) in which she employs the notion of an arational action. More generally, Rey (2001) constitutes a pointed and cogent critique of received views of the importance of the personal-subpersonal distinction.

the person (or subject, or self) is held responsible for them and is obligated to reinstate rational coherence.

In contrast to the natural world, the personal level provides the home for normative properties. The personal level is the distinctive realm of obligation, oughts, and shoulds. Accordingly, it is the realm of responsibility for one's choices and one's actions. The subject's reasoned deliberation causes her actions, and in the wake of such action, praise and blame are appropriate; someone who chooses to divert his or her employer's funds to a private bank account has made a poor decision and can rightly be blamed for it. Consider the contrasting situation in the natural world. It is the realm of blind mechanism, deterministic forces, and natural laws. There is no ought or should in the natural world; there are simply the facts about what happens. Let's say, as Newton thought, that the force acting on a body equals the body's mass times its rate of acceleration. In that case, F = ma describes a widespread pattern of occurrences in nature, with no evaluative connotation. It is not as if nature *should* be that way or *ought* to be that way; it just is that way.

The subpersonal level is, roughly speaking, the portion of the natural realm pertaining specifically to the processes that produce action or behavior. Often, when drawing the distinction between the personal and subpersonal levels, the latter is associated with neural activity. But, the more general idea is that the subpersonal level contains whatever forces or processes produce human behavior by simple cause and effect, the kind of cause and effect associated with the natural sciences, where natural law and mere mechanism reign. This might be characterized as the engineering level (cf. Dennett 1987); persons, minds, consciousness, and selves appear at the personal level, while the subpersonal level houses the mechanical rigging, ropes, and wires—whatever makes possible, enables, accounts for, grounds, or implements the personal-level states and processes, whether it be computational processing of linguistic form or primary visual cortex's detection of edges.

3. Cognitive Science and the Personal-Subpersonal Distinction

How are cognitive science and the personal and subpersonal levels related to each other? According to the Received View, personal-level facts are identified independently of the scientific endeavor; they are known *a priori*, by introspection, by conceptual analysis, or by common sense. Cognitive science then investigates the subpersonal mechanisms or processes that make such personal processes possible (or actual). On this view, cognitive science works in the service of a certain kind of philosophy—introspection, conceptual analysis, and reflection on the deliverances of common sense.

Consider examples of supposed personal-level phenomena: someone's knowing English; someone's being addicted to cocaine; someone's being in pain; someone's being able to reason well; someone's being in love with someone else; someone's being a skilled chess player; and someone's visually perceiving the layout of his or her immediate physical environment. These are, it is thought, states or capacities of the entire person. You, as an entire human being or person, can, for example, navigate a room visually. Cognitive science is responsible for explaining how stimulation of your retinal cells by photons gets processed in successive stages, from activity in bipolar cells, on through the lateral geniculate nucleus, to occipital cortex, and so on, resulting in, or otherwise facilitating, a person's having a perceptual experience that can guide her smoothly across the room.

The Received View prevails among a large number of influential philosophers of mind—including John McDowell (1994), Tyler Burge (2010), Susan Hurley (1998), Mark Rowlands (1997), Jennifer Hornsby (2000), Martin Davies (2000*a*, 2000*b*), Jose Luis Bermúdez (2000), and Nicholas Shea (2013), among many others. They take non-experimentally produced personal-level phenomena as the (primary) *explananda* of cognitive science or to constrain cognitive scientific modeling (model choice, in particular) and theory construction in some other significant way. To be clear, various among these authors allow more or less revision of *a priori* or commonsense commitments regarding personal-level phenomena. Some significant constraint from the personal level remains in place, however, even on this weaker view; for even if characterizations of supposed personal-level phenomena are allowed to be defeasible, they provide a presumptive constraint on cognitive-scientific modeling and theorizing, in the way *explananda* do; they provide targets of explanation and therefore put regulative or evaluative pressure on cognitive science, in the way that one would think only data or patterns in data would. And, of paramount importance, they are thought to do so without actually being data of the standard sort that ground experimental sciences, that is, measurable, replicable results. Cognitive science is, on this view, engaged only (or primarily) in 'vertical explanation' (Drayson 2012; cf. Bermúdez 2000, Colombo 2013, Cummins 1983), at least to the extent that cognitive science appeals to the subpersonal in formulating explanations; cognitive science does its proper job when it takes an independently established mental state, ability, or capacity as target and investigates the mechanical, computational, or neural processes that enable (McDowell 1994) that personal-level state, ability, or capacity.

A philosopher of science should suspect that something has gone wrong with this interpretation of cognitive science; for, in no other science do pre-theoretical intuitions about phenomena constrain science in the way that *a priori* claims about the personal level are thought by many philosophers of mind to constrain cognitive science. This is not to say that intuition or guesswork play no role in structuring a science. Sciences begin with hunches about what sort of observations are of a piece with each other, that is, what sorts of observable phenomena fall in that science's distinctive domain. But, these amount to hunches about what range of data will be accounted for by similar models, which hunches are hostage primarily to the further collection of, and modeling of, data. It's not a matter of there being higher-level phenomenon about which commonsense facts must be respected when science does its work. Our best evidence that two systems are both harmonic oscillators requires that our best models of the behavior of those systems share a deep theoretical unity (e.g., equations used to model them have the same basic form). Physics models the behavior of the systems, and the lessons about those systems—whether they're of the same kind and so on—emerge from an interpretation of those models; it's not a matter of being committed to the construction of a model of the implementation, ground, or enabling conditions of some higher-level phenomenon—the property of being an oscillator—as it is conceived of by the folk or characterized by conceptual analysis.

To give the reader a taste of the Received View, and its shortcomings, consider the following passage from Tyler Burge's *The Origins of Objectivity*:

> The science of perceptual psychology is motivated by the goal of contributing to an explanation of how individuals perceive. More particularly, vision science assumes that individuals have approximately accurate visual perception some of the time. And

> it tries to contribute to an explanation of how such perception comes about to the extent that it does. (Burge 2010, pp. 87–88)

This should strike philosophers of science and cognitive scientists as strange. Vision science is out to develop the best models of the most judiciously collected, replicable data—for instance, of verbal reports of the shape or orientation of stimulus items. Of course, any philosopher of cognitive science or cognitive scientist worth their salt should want to identify theoretically interesting patterns in such models. And, at any given time, these more theory-oriented efforts at generalization over models might paint a picture of vision according to which most percepts are approximately accurate; *but such efforts might not*, depending on the state of the science at that time. It is likely that, in the end (if ever there be one), the correct answer will have emerged from our best models of vast amounts of data; but a legitimate science does not take such answers as given prior to enquiry. Everyday ideas concerning the mind are likely to influence and inspire experimental design in cognitive science, but that's quite different from establishing the goals—the very standards of success—for cognitive science.

More to the present point, there's no reason to think that vision scientists are committed to personal-level states (note that, on Burge's preferred usage, 'individual level' plays the role of 'personal level'). Vision scientists do indeed care about the number of subjects involved in an experiment, in the standard parlance of methodology sections ("there were *n* subjects..."), and the reidentification of those subjects over time; but this practice carries no metaphysical commitment beyond what's necessary to index any particular piece of data to the organism that produced it. Such indexing is a matter of tracking the performance of a particular organism over time; it does not presuppose the existence of a higher level of reality or distinctive domain of personal-level properties.

Burge goes on to assert an even more striking form of philosophical priority in cognitive science:

> But, necessarily and constitutively, individuals perceive. Perceptual states, as distinct from transformations by which they are formed, are the individual's. Individuals perceive as a result of perceptual states' being formed in their perceptual sys-

> tems. Perceptual states are realizations of individuals' capacities. I think that this claim is a priori. (*Ibid.*, p. 369)

According to Burge, there's an entity, the individual, and, when visual perception occurs, that individual is in the state, or has the property, of perceiving; this is "necessarily" and "constitutively" true. Moreover, Burge claims to know this *a priori*, that is, that his belief is justified independent of empirical investigation. The strength of such claims is, however, deeply puzzling: perhaps perceiving a scene is a state of the individual; but perhaps it isn't.[2] The proof is in the pudding, and thus the pudding has to be made and inspected before one attempts to assemble such proof.

Of course, a philosopher who does not intend to integrate his or her philosophy of mind with contemporary scientific findings and who does not claim to draw any support for his or her views from the cognitive sciences might feel free to make pronouncements about the matter in whatever way he or she likes. Burge, however, is no such philosopher. His writings on vision are empirically sophisticated. He has clearly mastered a wide swath of vision science and intends to take it seriously. In the work quoted above, he surveys a wealth of results from vision science. Why, then, the insensitivity to the structure of scientific enquiry? Why the proclamation of *a priori* knowledge about personal- or individual-level facts—knowledge of facts that set the very standard of success for vision science? Why not argue, instead, that our best cognitive science makes positive use of the concept of an individual or personal level, acknowledging that it just as well might not have and that cognitive science might take a different turn in the future?

[2] One might stipulate that 'percept' can be applied only to states of the entire individual, but doing so would simply move the bump in the rug. Such stipulation introduces the possibility that humans do not ever perceive but instead have processes very much like perception, though these processes must be called by a different name ('schmerception' or 'perception*' or what have you), because they're not in any deep or distinctive way states of the individual. Bear in mind that, from a grammatical standpoint, states of individuals are easy to come by. Consider an organism in which bipolar cells are firing. One can describe that situation by saying, for example, "The bipolar cells in Joe's visual system are firing," which has the grammatical form of an attribution of a property to a subpersonal mechanism in Joe. But, one could just as well describe the situation as "Joe's being in the state of having his bipolar cells firing," which has the grammatical form of an attribution of a property (via the attribution of a state) to Joe as a whole. Thus, grammar alone is no guide to the metaphysics of levels in such cases.

Burge is no outlier here, at least not with respect to the general picture of the relation between cognitive science and philosophy of mind. In a relatively recent book chapter, Nicholas Shea runs through the details of some leading neural models of decision-making. And, summarizing the role of the experiments vis-a-vis the personal and subpersonal levels, he says:

> But it is unsatisfactory to assimilate the whole pattern of behavior to the subpersonal level. The subjects are fully conscious normal adults, behaving as they do in the experiment because they have understood and are following instructions... Subjects are motivated by the cash rewards available in the experiment and their behaviour shows sensitivity to the structure of those rewards. It is hard to deny that they are acting voluntarily when they select one stimulus over another....the thing to be explained—the subject's behaviour—is a voluntary action at the personal level. So the temporal difference model...provides a putative subpersonal level information-processing explanation of a personal level phenomenon. (Shea 2013, 1074)

As Shea sees things, cognitive science is "uncovering the constitutive basis of personal-level phenomena like believing, desiring, and perceiving" (*ibid.*, 1068). Thus, the Received View is, I submit, standard fare in philosophy of mind, among empirically sophisticated philosophers of mind, no less. Known facts at the personal level are treated as the phenomena, and cognitive science provides an explanation of them by identifying their subpersonal basis.

I contend that such authors as Burge and Shea present an erroneous picture of cognitive science, regardless of whether their marking of a personal-subpersonal distinction is useful for purposes unrelated to the understanding of cognitive science. Moreover, to the extent that they present a misguided philosophy of cognitive science, it is also a misguided naturalistic philosophy of mind. If one would like to understand the self in the age of cognitive science, the Received View threatens to mislead us, by promoting a view of cognitive science according to which the identification of the self is independent of the nitty-gritty work of the science itself.

In contrast to the Received View, consider the perspective of two of the founding figures of cognitive science, Alan Newell and

Herbert Simon, in the first major presentation of their groundbreaking research program. They say, "What questions should a theory of problem solving answer? First, it should predict the performance of a problem solver handling specified tasks" (Newell, Shaw, and Simon 1958, 151). On this view, personal-level capacities or abilities are not the constraining *explananda* of cognitive science. Rather the goal is to model observable, measurable data. (This is to be expected; cognitive science was meant to be a science after all.) And, the essential goal has remained the same over the decades of development in cognitive science: to model human performance (or the measurable aspects of the performance of nonhuman subjects). Here we have two modern giants in cognitive neuroscience, Larry Squire and Eric Kandel, describing the cognitive science of memory: "Memory promises to be the first mental faculty to be understandable in a language that makes a bridge from molecules to mind, that is, from molecules to cells, to brain systems, and to behavior" (2000, 3). What is it, according to Squire and Kandel, to cross the bridge from molecules to mind? It is decidedly not the preservation of commitments to antecedently identified personal-level states. The goal is to model behavior of the relevant sort by appeal to the tools of natural science, from biology to systems-level analysis neural activity.

This is not to say that our everyday conception of human capacities has no role to play in cognitive science. It has many roles; it inspires experimental design, for example, and provides useful language for summarizing patterns of results—in introduction and discussion sections of scientific publications as well as in science writing. This is par for the course in the history of science, having nothing particularly to do with psychology or cognitive science. In the initial stages of the investigation of some phenomena, research is guided by pre-theoretic observations, which inspire and structure the investigation. But, that is quite different from the case in which the pre-theoretic observations or intuitions provide a hard constraint on the interpretation of the science, as it matures, or in which they set the standard of success for the science in question. Pre-theoretic commitments to personal-level facts do not guide model selection in the substantive way required by the Received View; if the Received View were correct, then, in a situation in which an otherwise superior model fits poorly with the supposed personal-level facts than does an otherwise inferior model, cognitive scientists would reject the former in favor of the latter. But, so far as I can tell, this is not standard practice in cognitive science. We should perform modus tollens, then, rejecting the Received View and sidelining the personal level in our interpretation of the structure of cognitive science.

What, though, should we make of the use of 'memory faculty' in the passage quoted above from Squire and Kandel? The literature is rife with such uses: cognitive science is meant to account for such phenomena as the learning of spatial layouts, the acquisition of language, the ability to recognize faces, so on. Each of these cases might be understood as a case in which cognitive scientists want to account for personal-level phenomenon. But, the "accounting for" in question is too thin to offer solace to philosophical proponents of the Received View. The use of 'memory' and 'learning' is not a reification of a personal level, beyond a bet that there is a relatively unified model of a range of behavior that we would normally refer to as 'memory-related';[3] it does not express a commitment to knowledge acquired *a priori* (or by introspection, common sense, or conceptual analysis) of facts about that level, which then provide the explanatory targets, *qua* standards of success, for cognitive science. Rather, in these contexts, use of such talk serves as a tool for indicating patterns in the data or in the models of those data; or has an organizational aspect; or reflects pretheoretical guesses (encoded in everyday language) regarding which behavioral data sets are likely to be such that the best models of them are unified (Rupert 2013, Colombo 2013). But, that's a far cry from a situation in which accounting for some supposed personal-level facts provides the explanatory target of cognitive science, providing a *desideratum* that guides selection between competing models.[4]

[3] See Tulving (2000, 41) for doubts that memory is natural kind. I read these doubts in the following way. Cognitive scientists working on memory are willing to offer vertical explanations of system-level capacities, so long as those capacities, individually, amount to natural kinds, where our best evidence of such status is a certain unity to the modeling of the various data associated with that general capacity. Tulving sees significant diversity in the modeling of such data and thus resists the idea that there is a general, system-level capacity, *memory*, to be explained vertically. Note, however, that even in cases in which the system-level capacity is accounted for vertically, commitment to its existence is a contingent matter depending on the success and unity of certain modeling strategies; moreover, the establishing of such a system-level capacity does not establish the existence of a personal level of reality at which appear the distinctive properties that the Received View associates with the personal level.

[4] The comments in the main text might seem to presuppose too narrow a concepttion of cognitive science, as being concerned only with models of the mechanisms that produce behavior (in the generic sense of 'mechanism'). The reader might reasonably suspect, then, that this narrow conception of cognitive

As a further example, consider one of the most currently influential "top-down" research programs in cognitive science, Bayesian cognitive modeling. Among leading proponents of this view are Thomas Griffith and Joshua Tenenbaum. Their work shows less concern for mechanism than it does for highly abstracted, formal models, giving at least initial priority to the elements of the Bayesian frameworks: hypothesis evaluation and probabilistic inference. This might sound as if it shares the top-down perspective of the Received View. After all, it does seem natural to think of evaluating hypotheses and performing inferences as personal-level matters. But, notice the way in which they and co-authors characterize such constructs: "Hypotheses can take any form...as long as they specify a probability distribution over observable data. Likewise, different inductive biases can be captured by assuming different prior distributions over hypotheses" (Griffiths et al. 2010, 358). The sort of hypothesizing and probabilistic inference in question bears almost none of the supposed marks of the personal level. The hypotheses are not consciously entertained; the typical

science stacks the deck against personal-level approaches. Two remarks in response: First, an emphasis on the modeling of the mechanistic processes that produce behavioral data is what makes cognitive science a distinctive enterprise (differing in character from, say, the mere formal modeling of economic rationality or the mere collection of data in behavioral psychology labs—not that such projects have no role in cognitive science; they inspire the search for computationally or neurally realistic mechanistic models). Second, and this is the deeper point, the integrative project gives cognitive science its distinctive appeal as the ultimate "science of the mind"; only a science that integrates its formal models with, and accounts for behavioral data by appeal to, biologically plausible mechanistic models, satisfies the fundamental scientific urge to understand how all of nature, including minds, hangs together. In other words, there are excellent reasons why cognitive science, as conceived of here, emerged and flourished. Early modelers saw the promise of accounting for the forms of behavior that seemed particularly intelligent (language use, theorem proving, etc.) within an integrated scientific conception of nature, by using models of internal mechanisms that allow for precise prediction of measurable behavior (Newell, Shaw, and Simon 1958, 152, 155–156; see also Simon and Newell's more explicit, retrospective description [1971, 147–148] of what they were up to in their early work). This holds despite the fact that, early work in cognitive science tended to focus on information-processing models rather than on neural modeling. Early modelers were sanguine about the ultimate grounding of their proposals in the brain (Newell, Shaw, and Simon 1958, 163), partly because the place of computational processing in the physical realm was well-enough understood; computing machines could, after all, be built from physical parts performing operations that, demonstrably, human brains could perform (McCulloch and Pitts 1943). Thus, I do not rely on too narrow a vision of cognitive science; its attempt to model the mechanistic processes that produce measurable human behavior provided the impetus for its early development, and the extent of its success in this regard explains its rise to prominence and remains central to its status as the going science of the mind.

human cognizer is not specifying probability distributions over observable data. The inferences are not of the slow, deliberate kind. These constructs are far removed from the folk framework; neither Bayes's Theorem nor the axioms of probability theory from which it can be derived are part of folk wisdom or folk psychology. The process of updating probability distributions in response to data during, say, language processing (or any of the other examples that Griffiths et al. discuss) is not available to introspection; it does not guide verbal report. There's no reason to think the formation of hypotheses of this sort must be done by the whole person, as opposed to being done by components of her cognitive system.

If the Bayesian approach overlaps with personal-level theorizing at all, it might be in the rational nature of probabilistic inference, which many proponents of personal-level theorizing might identify as normatively laden. Nevertheless, the sort of normativity in question—there being a correct way to derive posterior probabilities from prior probabilities given some new data—hardly qualifies the top-down deployment of Bayesian models as a kind of personal-level modeling that friends of the Received View would embrace as their own. First, as noted above, the states and processes in question have very few, if any, of the commonly cited personal-level characteristics. Second, Griffiths et al. hope ultimately to contribute to cognitive science's search for models of human performance: "Although cognitive modeling and machine learning are two different enterprises, a basic challenge for both is to match human-level performance in domains such as language, vision, and reasoning" (*ibid.*, 363). In keeping with the conception of cognitive science that I stress throughout, the goal of cognitive science—whether, methodologically speaking, its investigations proceed in a top-down or bottom-up fashion—is to model the data successfully, not to respect and vindicate prior commitments to claims about an ontologically distinctive personal level. Third, on the vision of Griffiths et al., Bayesian cognitive modelers match human performance by guiding the search for mechanisms that perform Bayesian inference "in a variety of implicit and approximate ways" (*ibid.*, 362). More generally speaking, Griffiths et al. hope for a "synthesis with more bottom-up, mechanistically constrained approaches to modeling the mind" (*ibid.*, 362). On this view, "Probabilistic models are a tool for exploring different sets of assumptions about representations and inductive biases, making it possible for data to lead us to an account of human cognition" (*ibid.*, 363). Bayesian models do not reflect supposedly *a priori* truths about a personal level, truths that then strongly constrain the selection of models in cognitive science (leading cognitive scientists to reject otherwise superior models of the data because they don't portray human cognition as optimally Bayesian). Rather, the use of Bayesian principles helps us to formulate models of the actual human cognitive process and understand why that process might be useful to the organism, even if actual human cognition is not exactly Bayesian. Fourth, and closely related to the preceding point, for Griffiths et al., the top-down approach is an empirical bet about methodology. They offer empirical arguments for their approach, for example, that top-down Bayesian cognitive modeling is more likely to explore the space of possibilities effectively, and less likely to get bogged down in dead-ends, than is a mechanisms-first approach (*ibid.*, 358). Implicit in this style of argument is a commitment to contingency: if the competing mechanisms-first approach ultimately produces models that account for human behavior, while the top-down Bayesian approach flounders, Griffiths et al. will have lost their empirical bet. They do not take Bayesian principles to be "necessarily and constitutively" true of human cognition.

I am not arguing for eliminativism about the kinds of states appealed to by commonsense psychology (cf. Churchland 1981). Absolutely not. Perhaps such states exist; and if they do exist, they might be much as the folk conceived of them; but, then, again, they might not be. I take no stand on any of these issues. Rather, my point is that, if cognitive science is to vindicate such things as beliefs, intentions, and conscious states, such states had better show up in the models of the relevant behavioral data or of other relevant sorts of third-person data, such as imaging data that is meant to help guide the modeling of behavioral data.

And similar remarks apply to the distinction between a personal and a subpersonal level. Models in cognitive science incorporate whatever elements prove to be useful—place-indicating grids of hippocampal cells, decay-rate parameters for items in a visual buffer, arrays of on-center-off-surround cells, operations of categorization by a feature-matching algorithm, etc. Typically, however, these models account for the data directly; they do not run through personal-level constructs. Subjects (in the generic, organismic sense) exhibit patterns of responses. Cognitive scientists spend their time recording and modeling those responses, as well as collecting data on the neural processes involved, which will also be modeled as part of the data set that includes the record of the behavior of interest that was exhibited while imaging was being done. Such modeling could, in principle, vindicate the distinction between the personal and subpersonal levels, but that would require qualitative differences to appear in a wide range of models—differences that do some causal or explanatory work in the models—such that those qualitative differences map reasonably well onto the distinction between the two families of properties of the sort discussed in section 2, above. At present, it is difficult to find such a distinction in extant cognitive-scientific models.

Philosophers of mind seem to want to subvert or avoid this way of approaching cognitive science, by building the personal-subpersonal distinction into the structure of cognitive science from the outset; that seems to be the point of setting up supposed personal-level facts as the anchor, constraint, and *explananda* of cognitive science. But, so far as I can tell, that is bad philosophy of science. It would seem to be a molding of the structure of cognitive science to fit one's philosophical commitments arrived at by a different route.

Of course, philosophers of mind interested in cognitive science might pursue other goals than to articulate the structure of cognitive science. They might, for example, ask whether cognitive science, understood on its own terms, can inspire new uses of, or indirectly reveal new contours of, folk concepts, regardless of whether cognitive science vindicates the folk states or properties. A philosopher doing conceptual analysis might, by analogy, find that scientific results inspire the construction of a thought experiment, for example. Or, working by analogy, a philosopher proposing an account of, for instance, the self, belief, or weakness of will might pattern her account after some aspect of the structure of a cognitive-scientific model or family of models. But, it is absolutely essential not to conflate the fruitfulness of these relatively innocent ways of approaching cognitive science philosophically with support for the Received View. On the Received View, cognitive science answers questions about the self only by finding out how the self—which we know to exist on the personal level and which does not necessarily reflect an aspect of cognitive-scientific modeling—is implemented at the subpersonal level. Thus, on the Received View, if one would like to know what cognitive science tells us about the nature of the self, the answer is essentially "nothing" (Hornsby 2000); or, at the most, it might supplement or revise our personal-level conception of the self (Davies 2000*a*, 2000*b*, Bermúdez 2000, Colombo 2013). But, even the latter, weaker positions maintain the core commitments of the Received View; they accept the existence of a distinct personal level, the utility of a great amount of personal-level explanation, and the role of personal-level commitments as constraints on the cognitive-scientific enterprise. Once, however, we have set aside the Received View, we can query cognitive science directly: "Is there something self-like in cognitive-scientific models?" Here I think the answer is probably 'yes', a matter I return to in the closing section.

4. **To reiterate, then, I do not claim that the sorts of states associateed with the personal level—beliefs, desires, consciously experienced visual percepts—do not exist or that cognitive science eliminates them. Rather, my point is that if such states exist, we should expect to find them at the same level as mechanical, so-called subpersonal cognitive processing. I acknowledge the pos-sibility that a substantive personal-subpersonal distinction might emerge from cognitive science itself; but, regarding this possibility, I think the data speak against it at this point. Normative properties, for example, appear to do no work in extant models in cognitive science (cf. Drayson 2012) (though, to be clear, subjects' mental representations of normative properties do appear in extant mod-els—Klucharev et al. 2009, Klucharev et al. 2011). Cognitive-scientific models of conscious state do not place them at a distinctive level; the global workspace appears in models of consci-ousness right alongside the nonconscious processing sensory input (Dehaene, Changeux, and Naccache 2011). Thus, rejecting the existence of a personal level (or rejecting the scientific utility of personal-level explanation) *does not entail the rejection of beliefs, desires, etc.* The two issues are orthogonal—unless one holds the implaus-ible position that appearing at a higher ontological level from the level of, say, a computationally modeled language parser is a neces-sary condition on something's being a belief, desire, conscious state, etc.Empirical Support and Illustration**

This section describes empirical work that illustrates and reinforces both the negative and the positive messages of Section 3. To be clear, these messages are as follows:

> *Negative Messages.* Cognitive science does not provide evidence of the existence of a distinctive

> personal level and does not presuppose the existence of such a level as a structuring principle; thus, the Received View of the relation between cognitive science and questions in philosophy of mind should be rejected, and empirically oriented philosophers of mind should not expect to find the human self at the personal level;
>
> *Positive Messages.* Cognitive science deals in the modeling of objectively measured, replicable data; the relevant modeling practices leave room for the appearance of such states as belief, desire, and conscious experience alongside mechanisms, states, and processes normally associated with the subpersonal level;[5] cognitive science also leaves room for the appearance of a substantive personal-subpersonal distinction (though, as indicated in *Negative Messages*, it seems unlikely, in fact, to appear).

Let us look now in some detail at a paper by Lau, Rogers, and Passingham (2007), entitled "Manipulating the Experienced Onset of Intention after Action Execution." I focus on this paper for various reasons. The experiments to be discussed are of inherent interest and perhaps of particular interest to philosophers thinking about free will, action, and responsibility. Reader be forewarned, however. This inherent interest introduces a potential distraction. I do not represent these results as definitive and warn against the thought that my argument somehow rests on such a claim. Thus, two further reasons take pride of place. First, the kind of modeling that shows up in this paper manifests the norm in cognitive science. There is nothing idiosyncratic or off-beat about it. It is entirely representative, specifically with regard to the claims I make about the role of the personal level in cognitive-scientific modeling. Second, the article's title is chock full of philosophically loaded words: 'experience', 'intention', and 'action'. Experiences are normally taken to be conscious. Actions are normally taken to be

[5] Compare Gendler, "But alongside that belief there is something else going on" (2008, 635; and similar language is repeated on 636 and on 637), where that something else is what would normally be labeled 'subpersonal processing', although Gendler herself does not use the terms 'personal' and 'subpersonal'.

the expression of rationality, issuing from the subject's desires and beliefs that have combined to cause the formation of an intention to act. If one were ever to find a paper in cognitive science that incorporates, or is structured by commitments regarding, the personal-level, this paper would seem to be it. Nevertheless, the most parsimonious, least convoluted way of understanding Lau et al.'s model—that is, a straightforward reading of the model—reveals interaction between various same-level subpersonal states that produce the data collected, with personal-level states playing no role.

To frame the study, consider the possibility that one's conscious experience of forming an intention to act occurs after the neural processes that initiate the action in question. Experimental evidence of such a situation would be striking, for it would appear to show that one's conscious mind is not controlling action, contrary to intuitive reports, at least in some cases (Libet et al. 1983, Libet 1985).

It is, however, notoriously difficult to record absolute timing of the kind of events in question (see Dennett 2003 for a discussion of some of the complications). Thus, Lau et al. design experiments that probe only relative timing. The thought is that if a subject's report of the timing of his or her conscious intention to act can be manipulated by a neural intervention at the time of, or shortly after, the execution of the action, then that intention would not seem to be fully formed until the time of, or even after, the execution of the action. After all, if it had been fully formed and operative in the production of the behavior, why would the subject's estimate of the time of occurrence of the intention change as the result of events that happen after the cause of the action?

Let us look more closely at the design of Lau et al.'s main experiment. Subjects rest their heads on a support and look at a timing display of the type shown in Figure 1.

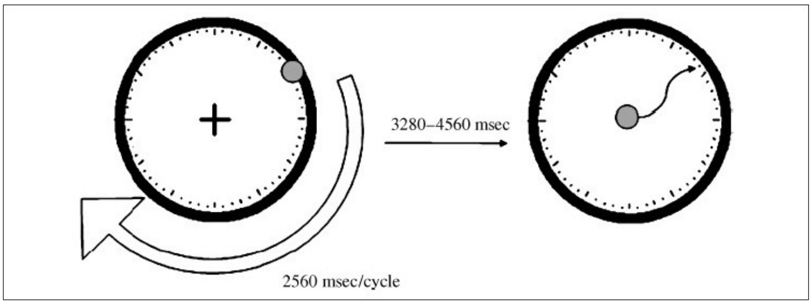

Figure 1. The timing device used by Lau, Rogers, and Passingham. Subjects attend to the face of the clock during the experiment. At the beginning of each trial, the dot moves from its center position to the perimeter and begins circl-

-ing clockwise, in the manner of a second hand sweeping round a clock face. At the end of each trial, the dot returns to its center position. The subject must then drag the dot, using a cursor in the subject's right hand, to the position on perimeter where the dot was at the time of the relevant experimental event (for example, when the subject first experienced a conscious intention to move his or her left index finger). Reprinted from Hakwan C. Lau, Robert D. Rogers, and Richard E. Passingham, "Manipulating the Experienced Onset of Intention after Action Execution," *Journal of Cognitive Neuroscience*, 19:1 (January, 2007), pp. 81–90. ©2007 by the Massachusetts Institute of Technology.

The subjects fixate on the dot in the middle of this fine-grained timing device, which begins in the center and then moves around the marked area on the circumference of the clock (with a total period of about 2.5 seconds). The relevant experimental events occur while the dot is moving. The subject is meant to index those events to the time of their occurrence, by reference to the clock. The subject does so by controlling a cursor with her or his right hand. The dot moves back to the center after the experimental events, and the subject grabs it with the cursor and moves it to the position that indicates the timing of the event about which the subject is being queried.

What are the events? The subject is told to press a button with her or his left index finger, at whim. Then, depending on the trial, the subject is asked to use the cursor to indicate either the time at which the conscious intention to move was formed or the time at which the action itself (the left index finger's pressing of the button) occurred. Lau et al. introduce four manipulations. They apply transcranial magnetic stimulation (TMS—which in essence scrambles the information encoded in whatever bit of cortex is targeted by the disturbance in the magnetic field caused by TMS), either at the exact time of or 200 ms after the action has been executed; or, they apply sham TMS (indistinguishable to the subjects from genuine TMS, except that no magnetic field is generated) at one of the two intervals in question. The actual TMS pulses target pre-supplementary motor area (pre-SMA), the part of the brain in which a wave of activity (the readiness potential) begins to build up prior to motor output.

Recall the logic of the experiment. If the subject had formed a conscious intention to move his or her left index finger, which then caused that action to happen, and, say, 200 milliseconds after the action, someone scrambled part of the subject's pre-SMA, the subject's report of the timing of the intention should not be shifted in time; such scrambling should be irrelevant. But, the results seem to conflict with that picture of the process. There is a main effect of genuine TMS v. sham TMS in both the case of reports of the timing of the intention to act and reports of the time of the action itself. (In each case—report of timing of action or report of timing of intention—there was no significant difference between the effect of administering TMS versus sham TMS at the time of the action or 200 ms after the action.) Moreover, the reported timing of the intention is on average shifted in one direction when subjects were queried about the time of action and in the other direction when subjects were queried about the time of the formation of the conscious intention to act. The reported timing of the intention was on average nine milliseconds or sixteen milliseconds earlier depending on whether the TMS was given a zero delay or 200-millisecond delay. The reported timing of the movement was either fourteen or nine milliseconds later in the TMS condition than in the sham TMS condition depending on whether it was a zero delay or 200-millisecond delay. This difference in the direction of the shift creates something of a puzzle. Why would TMS push the report of the timing of the intention to act in one temporal direction and the timing of the report of the actual movement in the other direction?

The results themselves are of interest, but what's more important for present purposes is Lau et al.'s proposed model of this peculiar aspect of the results. That model employs Bayesian cue integration. Technical details aside, the basic idea is that part of the brain is responsible for estimating the timing of events and does so by integrating signals arriving from various other parts of the brain during a temporally extended period or arriving from the same parts of the brain at different times in the relevant temporal window. And, each of those signals (that is, cues) is treated as having a different degree of reliability; it has a noise term associated with it that indicates, for example, the probability that the signal arriving from that source is not accurate. The part of the brain responsible for determining such timing can then cause motor output, in this case, the dragging of the mouse to a particular point indicating the estimated timing of the event in question.

A useful analogy is to one's asking a collection of different witnesses (or the same witness at different times) for testimony and then balancing the various reports against each other. Each witness will be treated as having a certain degree of reliability, which will be a factor in determining the best overall story to construct. In Lau et al.'s model, the degrees of noise associated with each of the different sources of information arriving at different times are analogous to levels of reliability assigned to various witnesses' reports.

Now consider the effect of TMS. Essentially, it reduces the reliability of the information in the area that has been hit with the electromagnetic pulse. That area of the brain has its information scrambled, so to speak, by complete polarization. As a result, the part of the brain estimating timing treats the signal coming from the scrambled part of the brain as less reliable relative to the other "witnesses" and discounts the information accordingly.

In all of the trials in Lau et al.'s main experiment TMS targets the same area, pre-SMA, but depending on what judgment is being made—the timing of the conscious intention to move or the timing of the movement itself—the contribution of the pre-SMA at the time of scrambling holds a different temporal position in the series of cues arriving in the part of the brain that estimates timing. When estimating the timing of the intention, events in the pre-SMA at the time of action or 200 ms after the action hold a late position among the "witnesses." Compare it, for instance, to the activity of pre-SMA, say, 200 ms prior to action, which will also make a significant contribution to the estimation of timing of the conscious attention, but that activity will be an "early" contributor. Thus, in the case of the timing of the intention, a late source of information is being discounted, with the effect that earlier sources are given more weight. Assuming that the part of the brain estimating timing uses the temporal order of the arrival of information as itself an image of the timing of events relevant to the calculation of the timing of the intention, the effect in the case of the estimation of the timing of the intention is to move it earlier in time toward the more heavily weighted earlier signals. The later witnesses—which given their very arrival time contribute cues that the event happened later—are being written off as unreliable. Thus, the reported timing of the intention is earlier in TMS condition versus the sham-TMS condition. Again, the estimated timing of the intention is a function of earlier and later activity in motor cortex. If you introduce noise into the later activity, it is discounted and earlier activity gets more weight. Earlier reporting witnesses are trusted more. That moves the report of the conscious intention earlier in time.

In contrast, TMS shifted the report of the action itself in the other direction, that is, significantly later in time compared to the reported timing of the action in the sham-TMS condition. Why? According to Lau et al.'s proposed model, it is because the report of the action itself is based on a different set of "witnesses," a different set of signals, within a different, and later, temporal window. The report of the action itself is a function of later activity in motor cortex together with confirmation of the executed action by proprioception and vision. Compare: When one thinks one has raised one's arm, that thought is partly confirmed by the proprioceptive information of one's arm being at a different location as well as one's seeing it in a different position, out of the corner of one's eye; these sources are in addition to information coming from the cortical source of the outgoing motor command to raise one's arm and kinesthetic information arriving internally from the nervous system indicating motion. In the experimental context, TMS on pre-SMA at the time of the action or 200 ms after introduces noise early in the temporal series of these sources of information, that is, earlier than such sources as visual and proprioceptive feedback. Given that the earlier "witnesses" are being discounted, the later ones—proprioception, for example—carry more weight than they normally would relative to the earlier ones. So, the part of the brain estimating the timing of the action assigns a later time stamp.

So far, personal-level properties and states make no appearance in the model. But, might one think of the behavioral report of timing as the output of personal-level process, that is, as a personal-level action. One could then tell a complex story about how supposed personal-level states—paying attention, being conscious of the instructions—interact with sub-personal level states—the ones appearing in Lau et al.'s model—to produce a personal-level action. That is a possible interpretation, but why bother? From the standpoint of the model, such a gloss is entirely gratuitous. It's not metaphysics being read off of the science; that would be one thing, and something worth attempting to do. Instead, it is superfluous metaphysics added to the model. After all, personal-level states don't appear in the model, and the elements of the model don't neatly fall into two categories that could plausibly be thought to mark a division between the personal and subpersonal. The model reveals only a collection of same-level states interacting to produce the data. The states that control motor report are part of a set of states that interact in complex ways with each other, some of those states assign degrees of uncertainty, some of them construct a model of temporal relations, some of them constitute the readiness potential, some produce finger movements, all at a single subpersonal level, and so on. That's all that the model contains.

One might object that the subpersonal states included in the model make sense only against the assumed backdrop of such personal-level states as the subject's paying attention and understanding the experimenter's instructions. But, the question to be asked in response is whether the cognitive-scientific models of those processes—of paying attention or understanding instructtions—invoke personal-level states. And, I submit that, in the relevant respects, such models are no different from the model already considered. Although Lau et al. do not themselves address the question of motor control of the cursor, they do make preliminary remarks about attention, providing evidence that they've correctly identified the fronto-parietal networks that constitute the attentional system that tracks activity in pre-SMA; such a model, based on activity in fronto-parietal circuits, does not include personal-level constructs. On this view, background assumptions about personal-level states and processes are likely to serve only as useful ladders in cognitive science, to be kicked away when models of those very processes are developed.

Let us now consider a second case, more briefly, that of the interaction of implicit and explicit states and their co-contribution to the production of behavior. There is a growing literature on the behavioral effects of what are often called 'implicit attitudes' (attitudes, because they have an evaluative dimension), with particular interest in socially relevant implicit bias, for example, bias in hiring. One widely discussed kind of case is that in which a subject professes egalitarian attitudes about race but, when asked to make snap judgments in an experimental context, seems to show bias against members of a racial "out-group" (see Brownstein and Saul 2016, vols 1 and 2).

A well-known version of this kind of experiment explores racial attitudes in the United States, typically involving white and blacks. In these experiments, subjects (both white, and to some degree, black subjects) seem to display racial bias against blacks on the implicit attitude test (IAT), even when the subjects avow egalitarian attitudes. In this version of the IAT, a monitor displays, serially, a mixture of positively valenced words, negatively valenced words, images of stereotypically black faces, and images of stereotypically white faces. Subjects are told to sort the images into two disjunctive categories, as quickly as they can. In one condition, the categories are "black-or-good" and "white-or-bad"; in a contrasting condition, the categories are "black-or-bad" and "white-or-good." The subjects sort the stimuli by pressing a button on the left or a button on the right, one assigned to each disjunctive category. They find the cate-gorization process more difficult in former case than in the latter. Their reaction times are significantly slower when one of the cate-gories combines black and good than when the category in question combines black and bad. In other words, the subjects seem to find it easier to fit negative words, as opposed to positive words, into the same category as black faces in a situation in which they must react quickly, without reflection or deliberation. The thought, then, is that subjects—even those who profess egalitarian attitudes—have negative implicit attitudes toward blacks. And, of great concern is that such negative implicit attitudes might be driving social interactions, from the creation of unhealthy conversational dynamics to the ways in which hiring committees make snap-decisions about the degree of promise a given résumé manifests (when, for example, it has a stereotypically black name associated with it as opposed to a stereotypically white name).

This conveys a sense of why the research on implicit attitudes is of such importance and has received so much attention. That being said, experiments involving race are part of a much larger research program investigating the role of implicit states, partly as a way of constructing a more accurate picture of human cognitive architecture. (Some of this literature is connected to the more general exploration of the distinction between System 1 processing and System 2 processing—see Evans and Frankish 2009.) Thus, one finds in the literature experimental results on other topics, such as cigarette smoking and junk food. Take a subject who claims to prefer healthy snacks over sugary snacks but who shows a preference for sugary snacks on an IAT, where the disjunctive categories into which stimuli are to be quickly sorted are, for example, "sugary-snacks-or-pleasant" and "fruit-or-unpleasant" (and vice versa, on other trials). Imagine that at the end of the session, as the subject is leaving, the experimenter gives the subject the opportunity to choose one item from a collection of snack foods and fruit as part of his or her compensation for participating in the experiment. Is a subject who shows a more positive attitude toward sugary snacks on the IAT more likely to choose a sugary snack over fruit, even if the subject avows a preference for healthy snacks? The answer would seem to be 'yes' (Perugini 2005). It is tempting to think of the architecture-related message of much of this work as follows: explicit attitudes, those that subjects report, are personal-level states, while implicit attitudes, those that drive fast, automatic, nondeliberative responses, are subpersonal states.

But, that would, I think, be to jump to an unwarranted conclusion. According Perugini, "The key message is that implicit and explicit attitudes can interact in influencing behavior" (Perugini 2005, 41), in some cases modulating the effects of the each other. Assume Perugini is correct. Still, his claim does not itself rule out the possibility that such interactions criss-cross ontological levels. Yet, although that possibility is left open, that's not the natural way to read Perugini's statistical model (see Figure 2).

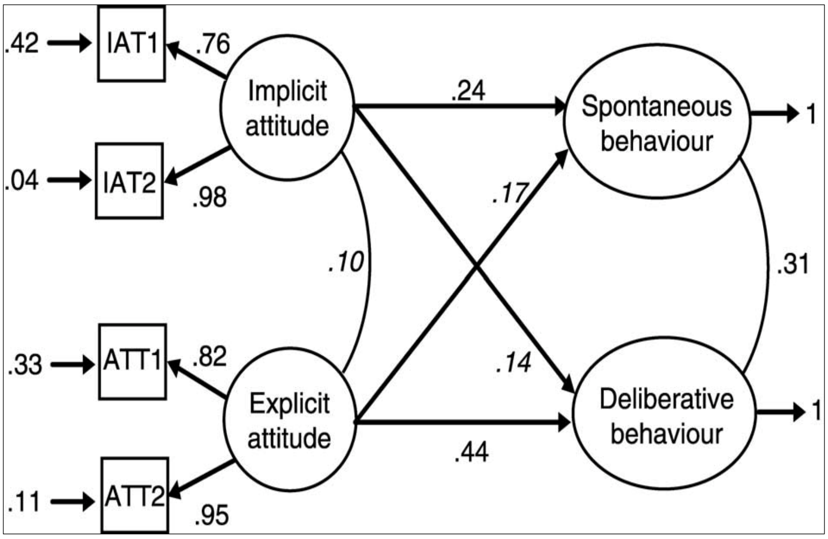

Figure 2. A graphical representation of various factors contributing to the production of deliberative or spontaneous behavior. Reprinted from Perugini, Marco. 2005. "Predictive Models of Implicit and Explicit Attitudes," *British Journal of Social Psychology* 44: 29–45. © 2005, The British Psychological Society.

Here we see a model that makes no use of an ontological distinction between a personal and a subpersonal level. There are explicit attitudes and implicit attitudes, each of which can contribute to the production of either kind of behavior: deliberative (normally associated with explicit attitudes) or spontaneous (normally associated with implicit attitudes). And, each of the two kinds of attitudes can modulate the contribution of the other.[6] No

[6] The italics in Figure 2 indicate that the values for interactive connections—those running between implicit and explicit attitudes—are nonsignificant. But, the analogous interactive terms take on significant values in the analysis of a structurally analogous experiment reported earlier in the same paper (an experiment on attitudes toward smoking and smoking behavior). Perugini gives no indication that he conceives any differently of the model of the factors relevant

element in the model corresponds to a distinctively personal level. There are only explicit attitudes and implicit attitudes, on the same level, correlated with results on various measures (IAT's and questionnaires), and interacting and co-contributing the production of spontaneous behavior and interacting and co-contributing to the production of deliberative behavior.

Perugini's conception of the architecture is not idiosyncratic. Bertram Gawronski and Galen Bodenhausen are leading figures in the field of implicit attitudes research and are the progenitors of the associative-propositional evaluative (APE) model of implicit attitudes. In a relatively recent paper, they say, "By making specific assumptions about mutual interactions between associative and propositional processes, the APE model implies a wide range of predictions about the conditions under which implicit and explicit evaluations show either converging or diverging patterns of responses" (Gawronski and Bodenhausen 2014, 188–189). This echoes the interactive view advocated for by Perugini, but in place of implicit attitudes, they talk about associative processes, and in place of explicit attitudes they put propositional processes, and in place of spontaneous behavior and deliberative behavior, they talk of implicit and explicit evaluations.[7]

What's the moral of the story, then, concerning cognitive architecture? One might claim that psychology is flat (as was claimed in Rupert 2015). That might be a bit of an overstatement. Perhaps some neural states implement other neural processes such that the subpersonal level itself subdivides into many levels (Churchland and Sejnowski 1992, Craver 2007). Nevertheless, psychology is, as one might say, "flattened from above." There is no distinct personal level that provides an external constraint on, by setting external standards of success for, cognitive science. On the "flattened from above" view, one is free to identify explicit attitudes with a distinctive kind of state, the kind of state that produces reflective verbal report, for example. But, the process of producing that verbal report appears side-by-side (as it were), not

to the results in the first case, even though that analysis is presented only discursively.

[7] They do so because they're inclined to think that there is only one underlying representation, but different ways of processing it, depending on the probe being used. If they are right about this, it would seem only to reinforce the negative point made in the main text, that an ontologically distinct personal level plays no role in the architecture.

above, many of what are typically identified as subpersonal states and processes.

5. In Search of the Cognitive-Scientific Self

Here is a different way to approach the idea of a flat psychology (or at least one flattened from above) that may help us to identify the cognitive-scientific self. Consider the distinguishing characteristics of various cognitive subsystems: the domain of application, typically understood as the kind of content the subsystem's representations carry, as well as the subsystem's contribution to the production of characteristic output (typically directed toward the objects, properties, or kinds in the domain in question). A face recognition subsystem is a face recognition system partly because it represents faces, and then allows access to that information in behavioral control, say, in the production of names that refer to individuals with the faces in question. A language parser is a language parser because it represents the syntactic structure and semantic content of incoming linguistic strings and it contributes to the control of, for example, verbal output when reading aloud. And so on, for visual processing, spatial navigation, and a variety of other subsystems that are normally placed at the subpersonal level. Nothing in my preceding arguments rules out the existence of one or more subsystems, appearing at the same level as these others, that is oriented toward what are typically thought of as personal-level phenomena (cf. the view taken in Drayson 2012 toward supposed personal-level states). One such subsystem may well contain information about the standing commitments of the very organism containing that subsystem.

Thus, we might identify the cognitive-scientific self with the subsystem responsible for encoding information—a biographical narrative, for example—particularly relevant to what is pretheoretically identified with the personal level. Dan Dennett has called this sort of construct the 'public relations department', because of its distinctive contribution to our explicit presentation of ourselves to conspecifics: "In fact, we wouldn't exist, as Selves 'inhabiting complicated machinery' as Wegner vividly puts it, if it weren't for the evolution of social interactions requiring each human animal to create within itself a subsystem designed for interacting with others" (Dennett 2003, 47). On this view, the self is a subsystem that manages interaction with others by providing socially felicitous reasons for our actions when queried, regardless of whether these accurately map onto the actual causes of our actions. But, Dennett's view is more expansive: "Wegner is right, then, to identify the Self that emerges in Libet's and his experiments as a sort of public-relations agent, a spokesperson instead of a boss, but these are extreme cases set up to isolate factors that are normally integrated, and we need not identify *ourselves* so closely with such a temporally isolated self" (*ibid.*, 48–49). And, Dennett goes on to describe ways in which this subsystem can also help to integrate one's own action over time, by keeping track of one's recently made decisions and the like. I agree that cognitive science is likely to yield a more expansive understanding of a self-representing, self-narrating subsystem. I say more presently about further functions of that representation presently. But, first, it's worth flagging a point of potential conflict. Although cognitive science may well detail many fruitful uses of a self-representation, cognitive science may yield a fruitful notion of a self as the cognitive system as a whole (Rupert 2009). In closing, I return to this possibility.

Assume there is one or more cognitive subsystems that encode and transform information to do with one's biography, experiences, recent decisions, etc. This may well come into play, as Dennett suggests, when we construct explanations of our behavior, to be offered to others. "I chose music school because I've loved music all my life," one might say when asked about one's enrollment in a music training program. This might or might not represent the decision-making process accurately; the person in question might actually have chosen to go to music school because, say, other motivation-oriented subsystems associate *being a musician* with *being attractive to potential partners*.

A standing representation[8] of one's own commitments, history, personality traits, and experience is likely to be useful for a variety of other purposes, however. For example, a self-representation may facilitate the search for and evaluation of options during decision-making. Exhaustive search is computationally or other-wise resource intensive. There's little reason to think humans engage in it, given limitations on the brain's computational power. Even systems with computational power far

[8] I use the term 'standing' quite loosely. The reconstructive nature of memory suggests that this information will itself be somewhat fluid. For example, the details and framing of one's life story may change somewhat from telling to telling—or, more generally, from instance of cognitive use to instance of cognitive use—without conscious awareness of such variation.

beyond a human's face search problems of the sort solved by computer scientists only by the use of heuristics (Russell and Norvig 2011). One commonly proposed way for humans to make decisions manageably is by satisficing, rather than optimizing (Simon 1956), that is, by hitting upon an option that is "good enough" rather than optimal. What, though, will count as good enough? Here one might find a role for the self-representation: an option that has made its way to the front of the queue might well be deemed "good enough," and thus selected, because of its sufficient degree of fit with one's self-representation. And, returning to the example of the preceding paragraph, this may be true even if other motivational states primed certain possible actions for consideration early in the search pro-cess; it might be that once music school was considered, it was chosen because of the high degree to which it meshes with one's personal narrative, even if a desire to be attractive to potential romantic partners, together with states automatically associated with that desire (say, images of musicians being treated as cool and desirable), caused the "music school" option to be pushed forward for consideration early in the process.

The availability of information—options or otherwise—might also be affected by one's self-conception. Which possibilities or facts "come to mind" in a given situation is determined by associative processing not available to consciousness. And, what drives these comings-to-mind may well, as much as anything, be degrees of associative strength between the representations in the subject's self-narrative and the information or possible actions in question. And, while this might have negative effects, such as confirmation bias, it might also serve to entrench further the contents of the self-representation, and thus enhance the positive effects of possessing a relatively stable self-representation.

Consider too the potential role a self-representation might play in the production of less deliberate action. Our more automatized actions—those taken while walking, driving, or holding a fluid, real-time conversation—may well have their course influenced by our standing representation of ourselves. In the case of walking, one's self-representation may help to determine gait; in driving, one's self-conception might affect how aggressively or cautiously one changes lanes; and in the case of conversation, how readily one makes critical remarks might be determined partly by one's self-representation.

This raises our final set of questions, however. Is there a well-demarcated portion of the cognitive system—the portion that keeps (or quickly reconstructs in context) a running biography—that should be thought of as the self? And, if so, should the self be identified with the mechanistic, neural, or computational portion of that system? Or, should the self be identified, rather, with the *content* of the representations in that system? Of course, further options move to the fore if the narrative-keeping subsystem is not well demarcated. The most obvious alternative perhaps is the entire cognitive system (the entire organism being another).

I tentatively opt for the cognitive system as a whole, while recognizing the fruitful purposes served by various processes that manage and deploy a record or a narrative concerning the organism's history, commitments, etc. My reasoning rests partly on the mutability of the contribution of such portions of the cognitive system (that is, on a concern that there is no discrete subsystem of the sort Dennett describes in the first quotation above), but even more so on the extent to which they collaborate and co-contribute, alongside other structures and subsystems at the same level, to the production of various forms of behavior that we treat as actions of the person. Beyond the sorts of examples given above, I have in mind such cases as that of the co-contribution of the body schema and the body image to the production of behavior (see Gallagher 2005 for a discussion of this distinction and the extent to which the operations of the body schema—normally thought to be part of the subpersonal cognitive system—suffuse the workings and outputs of consciousness—normally associated with personal-level phenomena).

What, then, is the cognitive system as a whole? As I see things (Rupert 2009, Rupert 2010), the cognitive system is an integrated collection of mechanisms that, in overlapping subsets, contributes to the production of a wide range of forms of intelligent behavior. What is intelligent behavior? Just as in any science, we must identify this, in the first instance, by example, that is, by pointing to various phenomena that we think are of a piece. Such examples might include engaging in conversation, giving correct answers on reading comprehension tests, creating line-drawings that guide the construction of a building having the same structure as the drawings, creating experimental apparatuses, giving lectures, writing books, playing complex games, and creating sculptures. In the architectural case, we observe behavior that's striking in its complexity and coordination. An architect makes line drawings. A general contract-

or carries those drawings away and subsequently contacts subcontractors, who contact laborers and suppliers, and so on. And, eventually, there comes into existence a large physical structure with the same form as the original drawings. This is incredibly impressive and cries out for explanation, as do a wide range of other forms of behavior that seem, broadly speaking, intelligent. Cognitive science bets that empirical investigation will yield a relatively unified set of models of at least most of these forms of behavior.[9]

What is built into the idea of overlapping subsets? The thought is that the human cognitive system contains various relatively specialized mechanisms each of which can contribute to the performance of a variety of tasks (see Anderson 2010, 2014, for a specific view about how the brain might support this arrangement). Take, for instance, a mechanism that identifies geometrical shapes (perhaps this appears in the word form area WF). It might contribute along with one set of mechanisms to the production of performance in reading aloud. It might contribute alongside a largely disjoint set of mechanisms in the solution of problems in geometry. It might contribute alongside yet another set of mechanisms while cleaning house. The idea, then, is that for any mechanism that is properly a part of the cognitive system, it contributes to a wide range of forms of intelligent behavior and does so by contributing in these shifting subsets of mechanisms. High degrees of overlap in such contributions integrate the various mechanisms into a single cognitive system; it is in virtue of all of a set of mechanisms having high degrees of overlap that they collectively constitute a single cognitive system.

How, on this view, should we think of subjects who profess egalitarian views about race but show conflicting results on the implicit attitude tests? Is such a person *really* racist or *really* egalitarian? The answer is "neither," in the sense that neither of these kinds of response represents the subject's true self. The true self is the entire collection of integrated mechanisms that produces various outputs under various circumstances. The subject is a person who produces *these* results when queried in a certain way and who produces *other* results when queried in a different way (cf. Dennett 1991). This, I submit, is the cognitive-scientific self, derived from

[9] Relatively unified, if cognition or intelligence are, in fact, natural kinds, but that remains to be seen.

the science itself by noticing the important role that a cognitive architecture plays in cognitive-scientific modeling. It is not a rationally coherent deliberator existing at a distinct level from cognitive-scientific models. It is a collection of relatively integrated mechanisms co-contributing to the production of intelligent behavior, and, although there may be central tendencies in the responses it produces, it is nevertheless a somewhat loosely knit team, the dynamics of the operation of which deserve the attention they continue to receive from cognitive scientists.

References

Anderson, Michael L. 2010. "Neural Reuse: A Fundamental Organizational Principle of the Brain." *Behavioral and Brain Sciences* 33: 245–313.

Anderson, Michael L. 2014. *After Phrenology: Neural Reuse and the Interactive Brain*. Cambridge, MA: MIT Press

Bermúdez, José Luis. 2000. "Personal and Sub-personal: A Difference without a Distinction." *Philosophical Explorations* 3 (1): 63–82

Brandom, Robert. 1998. "Insights and Blindspots of Reliabilism." *Monist* 81 (3): 371–392

Brownstein, Michael, and Jennifer Saul (eds.). 2016. *Implicit Bias and Philosophy*, Vols 1 and 2. Oxford: Oxford University Press

Burge, Tyler. 2010. *Origins of Objectivity*. Oxford: Oxford University Press

Churchland, Patricia S., and Terrence J. Sejnowski. 1992. *The Computational Brain.* Cambridge, MA: MIT Press

Churchland, Paul M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78: 67–90

Colombo, Matteo. 2013. "Constitutive Relevance and the Personal/Subpersonal Distinction." *Philosophical Psychology* 26 (4): 547–570

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press

Cummins, Robert C. 1983. *The Nature of Psychological Explanation*. Cambridge: MIT Press

Davies, Martin. 2000*a*. "Interaction without Reduction: The Relationship between Personal and Sub-Personal Levels of Description." *Mind & Society* 1: 87–105

Davies, Martin. 2000*b*. "Persons and Their Underpinnings." *Philosophical Explorations* 3 (1): 43–62

Dehaene, Stanislas, Jean-Pierre Changeux, and Lionel Naccache. 2011. "The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications." In S. Dehaene and Y. Christen (eds.), *Characterizing Consciousness: From Cognition to the Clinic?*, pp. 55–84. Berlin: Springer-Verlag

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

Dennett, Daniel C. 1991. *Consciousness Explained.* Boston, MA: Little, Brown and Company

Dennett, Daniel C. 2003. "The Self as a Responding—and Responsible—Artifact." *Annals of the New York Academy of Sciences* 1001: 39–50

Drayson, Zoe. 2012. "The Uses and Abuses of the Personal/Subpersonal Distinction." *Philosophical Perspectives* 26 (1):1–18

Evans, Jonathan St. B. T. and Keith Frankish (eds.). 2009. *In Two Minds: Dual Processes and Beyond.* Oxford: Oxford University Press.

Gallagher, Shaun. 2005. *How the Body Shapes the Mind.* New York: Oxford University Press

Gawronski, Bertram, and Galen V. Bodenhausen. 2014. "The Associative–Propositional Evaluation Model: Operating Principles and Operating Conditions of Evaluation." In J. W. Sherman, B. Gawronski, and Y. Trope (eds.), *Dual-Process Theories of the Social Mind*, pp. 188–203. New York: Guilford Press.

Gendler, Tamar. 2008. "Alief and Belief." *Journal of Philosophy* 105: 634–663

Griffiths, Thomas L., Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. 2010. "Probabilistic Models of Cognition: Exploring Representations and Inductive Biases." *Trends in Cognitive Sciences* 14: 357–364

Hornsby, Jennifer. 2000. "Personal and Sub-Personal: A Defence of Dennett's Early Distinction." *Philosophical Explorations* 3 (1): 6–24

Hurley, Susan L. 1998. "Vehicles, Contents, Conceptual Structure, and Externalism." *Analysis* 58 (1): 1–6

Hursthouse, Rosalind. 1991. "Arational Actions." *Journal of Philosophy* 88 (2): 57–68

Klucharev, Vasily, Kaisa Hytonen, Mark Rijpkema, Ale Smidts, and Guillen Fernandez. 2009. "Reinforcement Learning Signal Predicts Social Conformity." *Neuron* 61: 140–151

Klucharev, Vasily, Moniek A. M. Munneke, Ale Smidts, and Guillen Fernandez. 2011. "Downregulation of the Posterior Medial Frontal Cortex Prevents Social Conformity." *Journal of Neuroscience* 31(33): 11934–11940

Lau, Hakwan C., Robert D. Rogers, and Richard E. Passingham. 2007. "Manipulating the Experienced Onset of Intention after Action Execution." *Journal of Cognitive Neuroscience* 19 (1): 81–90

Libet, B. 1985. "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action." *Behavioral and Brain Sciences* 8: 529–566

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. 1983. "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act." *Brain* 106: 623–642

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–133.

McDowell, John. 1994. "The Content of Perceptual Experience." *The Philosophical Quarterly* 44 (175):190–205

Newell, Allen, J. C. Shaw, and Herbert A. Simon. 1958. "Elements of a Theory of Human Problem Solving." *Psychological Review* 65 (3): 151–166

Perugini, Marco. 2005. "Predictive Models of Implicit and Explicit Attitudes," *British Journal of Social Psychology* 44: 29–45

Rey, Georges. 2001. "Physicalism and Psychology: A Plea for a Substantive Philosophy of Mind." In C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents*. Cambridge: Cambridge University Press, pp. 99–128

Rupert, Robert D. 2009. *Cognitive Systems and the Extended Mind.* Oxford: Oxford University Press.

Rupert, Robert D. 2010. "Extended Cognition and the Priority of Cognitive Systems." *Cognitive Systems Research* 11: 343–356.

Rupert, Robert D. 2015. "Embodiment, Consciousness, and Neurophenomenology: Embodied Cognitive Science Puts the (First) Person in Its Place," *Journal of Consciousness Studies* 22: 148–180

Russell, Stuart, and Peter Norvig. 2011. *Artificial Intelligence: A Modern Approach.* Third Edition. Upper Saddle River, NJ: Pearson

Shea, Nicholas. 2013. "Neural Mechanisms of Decision-Making and the Personal Level." In Fulford, Davies, Gipps, Graham, Sadler, Stanghellini, and Thornton (eds.) *The Oxford Handbook of Philosophy and Psychiatry* (Oxford: Oxford University Press), pp. 1063–1082

Simon, Herbert A. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63 (2): 129–138

Simon, Herbert A., and Allen Newell. 1971. "Human Problem Solving: The State of the Theory in 1970." *American Psychologist* 26 (2): 145–59

Tulving, Endel. 2000. "Concepts of Memory." In E. Tulving and F. I. M. Craik (eds.), *The Oxford Handbook of Memory* (Oxford: Oxford University Press), pp. 33–43