The Self, Self-knowledge, and a Flattened Path to Self-improvement

Robert D. Rupert
University of Colorado, Boulder
November 14, 2020

I. Introduction

Philosophy of mind and philosophy of psychology aim to provide a clear and compelling account of the human self (Schechtman 2011); and epistemology hopes to illuminate self-knowledge, revealing its nature and identifying the conditions under which it can be acquired (Gertler 2010). Theories of these phenomena are, of course, not logically independent of each other. If one's theory of self-knowledge entails that humans acquire self-knowledge via method *M*, yet one's theory of the human self entails that the self is not the sort of thing that could be known about via *M*, then revision of at least one of those theories is in order.

This essay explores the connection between theories of the self and theories of self-knowledge, arguing (a) that empirical results strongly support a certain negative thesis about the self, a thesis about what the self *isn't*, and (b) that a more promising account of the self makes available unorthodox – but likely apt – ways of characterizing self-knowledge. Regarding (a), I argue that the human self does not appear at a personal level the autonomous (or quasi-autonomous) status of which might provide a natural home for a self that can be investigated reliably from the first-person perspective, independent of the empirical sciences. Regarding (b), I contend that the most promising alternative view of the self is revisionary: the self is to be identified with the cognitive system as a whole, the relatively integrated collection of mechanisms that produces intelligent behavior (Rupert 2009, 2010, 2019). The cognitive system teems with reliable, though not necessarily perfect, indicators (*cf.* Dretske 1988) of its own properties or of the properties of its proper parts, many of which are available for detection by, or the control of, further processes, such as motor control. I argue that indicating states should be treated as potential vehicles of self-knowledge, regardless of whether they are truth-evaluable states, such as beliefs. The investigation of self and self-knowledge frames discussion of a final topic, of some gravity: the way in which self-knowledge might contribute to self-improvement. In this regard, I emphasize the efficacy of certain forms of alignment between, on the one hand, elements of the cognitive systems corresponding to a more commonsense-based conception of the self and, on the other hand, processes associated with what is frequently referred to as 'implicit' cognitive processing (Evans and Frankish 2009).

II. Methodological Remarks

Much contemporary philosophical work pursues substantive knowledge – about justice, causation, reference, responsibility, mental states, rationality, free will, time, and much more – independent of, or with only passing sensitivity to, the work of the empirical sciences. In contrast, the present essay weights empirical results heavily. What justifies such an approach? It should go without saying that science has a *very* impressive track record; no other method of enquiry has cured polio, sent humans to the moon, produced personal computers, and so on (and on, and on) (Papineau 2001). Thus, when a philosopher weighs up various sources of evidence pertaining to a particular phenomenon, it only seems sensible that scientific results related to the matter at hand be given significant consideration. Philosophers who suggest otherwise seem to me to be like the coach of a sports team who refuses to play the team's star player.

Holism inspires the naturalism on offer, for it seems that only a certain kind of foundationalism – one that emphasizes indefeasible foundational knowledge and indefeasible inferences from those foundations – could justify the ignoring of scientific results (Quine 1969). For any topic or phenomenon of interest, most likely a panoply of evidence, issuing from a variety of sources, bears upon it. (And this is also true of the very question, “What bears on what?”) And, in most cases, the various bits of evidence do not point univocally to a single conclusion. Typically, then, one must sift through as much as evidence as one can manage on the topic at issue, in an attempt to figure out how best to interpret that body of evidence. A central aspect of this process involves the weighting of different sources of evidence. One transforms such holism, as an expansive attitude toward the inclusion of evidence, into naturalism by one’s choice of weightings, by weighting pieces of scientific evidence especially heavily relative to other forms of evidence.[1]

The preceding paragraphs hardly do justice to the richness and complexity of metaphilosophical debates, but it is hoped that they nevertheless give the reader some sense of why I do not think it is wrong-headed to engage deeply with the cognitive sciences. The remainder of the present section emphasizes ways in which scientific developments and other metaphilosophical insights bear on both (i) first-order questions concerning the self and self-knowledge and (ii) higher-order questions about the appropriate methods of investigating these phenomena. Even those readers left cold by the general methodological comments with which this section opened may nevertheless be moved by the content-specific considerations to follow.

The past sixty years of work in cognitive science paint human introspective powers in an unflattering light (Nisbett and Wilson 1977, Wilson 2003). This accumulated body of evidence recommends against the attempt to acquire self-knowledge by reflecting on one’s own mental activity. The individual subject may, in some circumstances, be especially likely to report accurately on what’s going on with her, but that alone does not support introspection-oriented

[1] “Doesn’t this put philosophers out of work?” the reader might ask. Not at all, no more than playing one’s star player puts other talented players on the team out of work. Readers of a certain age, who are fans of the National Basketball Association, might recall an oft-made sportscaster’s remark: “There’s no Michael Jordan without a Scottie Pippen.”

arm-chair investigations of the self; for, the empirical work in question calls deeply into doubt the subject's ability to remember accurately what she experienced. And, throughout a wide range of cases, the subject simply doesn't report accurately, even in the present moment, what's going on with her thought processes.

Perhaps, though, one can gain knowledge of the general phenomena of self and self-knowledge – knowledge about self-knowledge, as it were – simply by reflecting on the meanings or senses of 'self' and 'self-knowledge' or the concepts associated with these terms. Here, again, cognitive-scientific results recommend at least a moderate skepticism. For, not only is introspection not all that philosophers have tended to think; but neither would concepts, senses, or cognitively accessible meanings seem to be. Rather than being tidy sets of necessary and sufficient conditions – shared across subjects, accessible to reflection, playing the role of self-contained determinants of reference – concepts appear instead to be a motley of representational structures (Machery 2009). Moreover, even the tidiest view of conceptual analysis – as resolution into definitional components – presupposes primitives, at least if we reject definitional regress and circularity (Fodor 1998). What might such primitives be? On the assumption that a reduction of all concepts to sensory primitives is not in the offing, we face the possibility that the concepts of immediate interest, for example, SELF, are, as concrete cognitive units, primitive (even though they may be rich in associations that causally mediate their application). In which case, armchair analysis of meanings will do no good; the examination of genuine atoms reveals no internal structure.

It helps little in this context to argue that concepts or senses are only implicit, identifying them as whatever guides judgment about possible cases or as an amalgamation of, or abstraction from, such judgments (Chalmers 2004, 2012; for critical reactions, see Schroeter 2004a, 2004b, 2006, Rupert 2016). For this approach shares an apparently fatal flaw with approaches that recommend explicit analysis of concepts, as, for instance, sets of necessary and sufficient conditions: the problem of instantiation. It's one thing to produce an analysis of concept *C* or a pattern of judgments in one's application of *C* across possible cases. It is quite another to think that the analytic decomposition in question or the pattern of judgments in question accurately describes or reveals the nature of worldly properties or kinds – that is, the properties or kinds of things in the world that humans actually interact with and hope to understand. It's easy enough, for example, to have *a* concept associated with, say, 'self'. But, it's a further question whether that concept accurately describes or otherwise reflects (either explicitly or implicitly) the nature of actual human selves. A philosopher might contend, "Our concept of the self is a concept of something that meets conditions A, B, and C or drives pattern of judgment X." In response, someone might reasonably reply, "But, are *we* such things?" After all, concepts are cheap – they can be generated *ad infinitum* – and it is easy enough to associate a concept – either an explicit definition or a pattern of judgments about cases – with a word. Doing so hardly guarantees that the contours of that concept accurately reflect the nature of the thing to which the word in question is typically applied (or the nature of the thing that is causally responsible for current use of the word in question – Kripke 1980). This provides further reason to weight empirical research heavily. There's some sense in which humans can learn a lot about the self and self-knowledge by examining the relevant meanings or concepts; but for that process to be fruitful, we need to have settled on – among the infinite possibilities – the correct concept of self, the concept that corresponds to the property that we ourselves instantiate; in my view, effective identification of the right concepts, the ones that actually apply to things in our vicinity, must be guided at least partly by empirical work (Rupert 2016).

Another recent development should worry those who would give little weight to scientific evidence in their pursuit of an account of the self and self-knowledge. In recent years, philosophers of psychology and cognitive science have begun questioning the relevance of the commonly made, and commonly appealed to, distinction between the personal and subpersonal levels (Bermudez 2000, Rey 2001, Drayson 2012, 2014, Rupert 2011*a*, 2011*b*, 2013, 2015, 2018). This is no minor matter. Arguably, a commitment to this distinction – its metaphysical and epistemic importance – remains the most influential thread of Ryle's legacy (Rey 2001), as channeled through Dennett (1969). Its status as a dialectical staple is beyond question; it frames scores of debates and helps to define the rules of engagement throughout contemporary analytic and naturalistic philosophy of mind and epistemology (a comprehensive list of citations would fill pages, but here is a sampling: Hurley 1998, Rowlands 2009, 2010, McDowell 1994, Shea 2013, 2018, Miracchi 2017, Levy 2016, Davies 2000*a*, 2000*b*, Hornsby 2000, de Vignemont 2014, Lyons 2016, Ismael 2014, Fodor 1987, Robins 2017, Frankish 2016, Clark 2015, Holton 2016). For present purposes, the importance of the distinction is as follows: The supposed personal level – the level of reality at which conscious states appear, at which normative states appear, at which the states adverted to in folk psychological and rationalizing explanations appear – is thought *(a)* to provide the natural home for the self and its states and *(b)* to be a domain to which philosophers can gain access from the armchair – by introspection, folk-psychological common sense, *a priori* reflection, or conceptual analysis. And, notice that *(b)* has a methodologically powerful flip-side. It provides a highly flexible strategy for insulating philosophical claims from scientific results: although it might be allowed that cognitive-scientific results are of interest, they concern only the subpersonal level, and thus cannot provide counterexamples to claims about the personal level (Hornsby 2000). Invoking the epistemically (presumably grounded by the metaphysically) isolated nature of the personal level is used to justify the bracketing of philosophically inconvenient empirical results. But, reasons for believing in an epistemically (perhaps because metaphysically) autonomous personal level are rather thin on the ground. If one weights scientific results heavily, as I think one should, one will be struck by the lack of work done, in the sciences of the mind, by a substantive personal-subpersonal distinction, and one will thereby question the intuitions adduced in support of the applicability of a robust personal-subpersonal distinction to humans. To the extent that a comprehensive picture emerges from cognitive science regarding the architecture that produces human judgements and behavior, an architecture that does not presuppose or have built into it a robust distinction between the personal and subpersonal levels, one should doubt all the more strongly the philosophical intuition that such a distinction applies to the human case (regardless of how frequently 'I' and 'me' – and 'owner' and 'subject' – appear italicized in the philosophical literature).

Where, then, should the intrepid seeker of the self and self-knowledge turn amidst the modern, scientifically inspired malaise of hidden cognizing and veiled motivation – the 21st-century world of aliefs (Gendler 2008*a*, 2008*b*); of fast, automatic, subconscious System-One-style processing (Evans and Frankish 2009); of opaque operation of deep neural nets some of which might model fundamental workings of the cognitive self (Buckner 2018); and, to come at the issue from an angle familiar to those working in moral psychology, of malleable character (Doris 2002). Shall we despair of finding a path to self-improvement?

III. The Elimination of the Personal Level

A commonly held, bifurcated picture of human psychology places beliefs, desires, and states of conscious awareness at a so-called personal level, which is insulated from such processes as those assigning syntactic form to the incoming speech stream (Kim and Sikos 2011) or detecting the presence of blobs in early visual processing (Marr 1982), processes thought to appear at a "lower," subpersonal level (Drayson 2012, 2014, Shea 2013). What, however, is the personal level meant to be?

Dan Dennett introduces the personal-subpersonal distinction to mark differences between "levels of explanation" (1969, 90), between explanations that appeal to properly mental vocabulary – involving descriptions of what the *person* does – and explanations that are in the same vicinity but are physical and mechanical (*ibid.* 92–94). I say "in the same vicinity" because, for Dennett, when one moves from discourse about, for instance, the person's pulling away from the stove because the pain hurts to a discussion of "brains and events in the nervous system" (*ibid.* 93), one "abandon[s] the explanatory level of people and their sensations and activities and turn[s] to sub-personal level" (*ibid.*); and in doing so, "we abandon the subject matter of pains" and analyze "something else – the motions of human bodies or the organization of the nervous system" (*ibid.* 94). Dennett's focus on explanatory vocabularies allows him to legitimate personal-level explanations without holding that mental terms refer (*ibid.* 96). Setting aside the nuances of Dennett's program, philosophical use has, in the decades since Dennett introduced the personal-subpersonal distinction, tended more toward the material, rather than the formal, mode: some mental states are states of the subject *as such*; these, and only these, appear at the personal level (McDowell 1994).[2]

[2] One still encounters less metaphysical ways of talking about the distinction. Tyler Burge (2010), for instance, frequently talks about what is "attributable" to the subject (e.g., *ibid.* 95, 369), which has a more epistemological flavor to it, being about what some procedure of attribution would validate, rather than being about the thing itself to which the state or property is being attributed. Note, too, that throughout *Origins of Objectivity*, Burge shows a strong preference for the terminology 'individual level' and 'subindividual level' rather than 'personal' and 'subpersonal' levels. For the most part, the reader is left to infer (reasonably enough) the equivalence of these two distinctions (perhaps speculating that Burge chose the new terminology for stylistic reasons or for the purpose of precisification); but at various points Burge explicitly uses the language of the person-level and the subpersonal and treats this talk as equivalent to talk of the individual and subindividual levels (93, 369 n3).

Appeals to the personal-subpersonal distinction seem to have flourished alongside, and taken nourishment from, the development of a naturalistically oriented anti-reductionism in philosophy of science. Such anti-reductionism entails that, although physics constrains – metaphysically, from below – every special science, each (successful) special science nevertheless has its own integrity and autonomy, its own proprietary kinds and properties related by its own laws (Fodor 1974). Thus, it's no surprise that treatment of the personal level often mirrors the treatment of non-fundamental scientific levels – as containing a distinctive set of properties or processes, interconnected by a set of level-proprietary laws or models, and that can be investigated largely independently the levels below, typically thought to realize or otherwise metaphysically determine the higher-level facts, absent full-blown reduction (*cf.* Craver 2007, ch. 5).[3] The oddness of this situation is that, although Fodor may be right, in general, about the autonomy of the special sciences, the relevant special science – cognitive science – doesn't vindicate his own tendency to take the personal-subpersonal distinction on board. In fairness, Fodor's naturalism (at least sometimes) tempers the inclination to make *aprioristic* claims about

[3] It is natural to think of the personal level in metaphysical terms, as the "location" of consciousness, persons, and personal responsibility in the universe. In comparison, the biological level is metaphysically real, one might say, because properties appear at the biological level that don't appear in the catalogue of properties of interest to physicists, and, moreover, those biological properties – such properties as, say, patterns of predator-prey relations – are multiply realizable in systematically comprehensible ways and offer great causal-explanatory advantages when used exclusively. When one conceives of the supposed personal level metaphysically, on the model provided by anti-reductionism in philosophy of science, questions arise concerning which properties are distinctive of it and whether those properties have the sort of (quasi-)autonomous causal-explanatory power one expects to find in an independent science. One such set of properties is epistemological, although thought to appear partly because of the metaphysically distinctive nature of the level; on this view, whatever exactly the self is that emerges at the personal level, and whatever distinctive properties emerge at the personal level (e.g., responsibility, normative status, consciousness, capacity for performing actions in accordance with reasons), these properties give rise to a certain epistemic situation, the ability to gain knowledge about one's self directly – by reflection, introspection, or immediate self-awareness. One might argue for a more purely epistemological reading of the personal-subpersonal distinction, one that insulates its deployment from the criticisms developed here, by sticking closer to Dennett's original intentions (and closer to a typical anti-realist interpretation of anti-reductionism in philosophy of science). This strategy doesn't seem promising, for one upshot of the criticisms developed here is Dennett's personal-level explanatory vocabulary does not possess the power one would expect of an autonomous explanatory vocabulary worth wanting; it does not both *(a)* have sufficient autonomous explanatory value (in the way biology does), while *(b)* also dovetailing in the way autonomous sciences do with neighboring sciences (in the way, for instance, cellular biology and molecular chemistry do).

the personal level. Rather, he is impressed by the success of folk psychological predictions and explanations, and thus looks to cognitive science (or computational psychology) to vindicate such practice and its posits, including, it would seem, its commitment to the existence of personal-level states (Fodor 1987, chapter 1). Where Fodor seems to go wrong is in thinking that successful interaction with our peers – anticipating their behavior, interpreting their meaning, coordinating our actions with theirs – depends on our assuming that the states attributed to our peers populate a distinct personal level; or in his thinking that, if folk practice is driven by such an assumption, the assumption's accuracy helps to account for the success of folk practices and will be vindicated by cognitive science, alongside the vindication of beliefs and desires (regardless of whether they appear at a distinct personal level).

This is a tricky matter, however, from an exegetical standpoint. When Fodor equates the personal level with the domain of folk psychological (or belief-desire) explanation, he seems to want to vindicate it. Elsewhere, though, he seems more inclined to identify the personal level with the domain of conscious states or to treat it as a forensic matter, and he expresses doubt that scientific psychology has much use for the distinction between the personal and subpersonal, suggesting that, since the kind *conscious state* is not likely to be of causal-explanatory use to cognitive science, the construct of the personal level will not be either (Fodor 1975, 52).

Fodor exegesis aside, we should be clear about the logic of the relation between the autonomy of psychology (or cognitive science) and the existence of a personal level. One would expect that, if there is a personal level, an autonomous psychology will emerge, consistent with anti-reductionism in philosophy of science. But, the conditional does not run in the opposite direction: neither the existence of an autonomous psychology nor an autonomous cognitive science entails the existence of a personal level; science vindicates the existence of a personal level only if the personal-subpersonal distinction proves to be of causal-explanatory use to an autonomous scientific psychology or cognitive science. In short, if there is a personal level, expect there to be autonomous psychological science; but if there is an autonomous psychological science, that alone is no particular reason to expect vindication of a personal level (it depends on the details of the science). In what follows, I argue that the relevant science does not vindicate the commitment to a personal level, as that level is normally understood.

But, first, can any more be said about what the personal level is meant to be? Regrettably, the personal-subpersonal distinction has been marked out in more ways than can be manageably reviewed here. Sometimes emphasis is placed on the entire organism or individual in contrast to its parts; it is John who sees the tree, not John's occipital cortex (McDowell 1994, Lyons 2016). Often consciousness serves as the mark of the personal level; sometimes it's rationalizing explanation (Davies 2000*a*, 88–90, 2000*b*, 46, Shea 2013, 1064–1065, Frankish 2009, 90–91).

Essential for present purposes is the path that a personal level might open for a certain kind of account of self-knowledge: the idea, in particular, that the development of a theory of self-knowledge might draw its support from or presuppose an epistemically isolated domain, where the self resides and to which philosophers have some sort of (typically) nonscientific or distinctively first-personal access. Consider, for example, the idea that the autonomous personal level that is the level at which conscious states appear. If this is treated as a genuine *level*, and levels are, by definition, epistemically autonomous domains (they can be fruitfully investigated with little or no regard for lower levels), then one has a recipe for the construction of an introspection-based theory of self-knowledge. Conscious states are transparent to themselves, a distinctive property of them that partly warrants the claim that the realm in which they appear amounts to a distinctive level. On the assumption that conscious states constitute the self, to know oneself is no more or no less than to be aware of one's own conscious states.

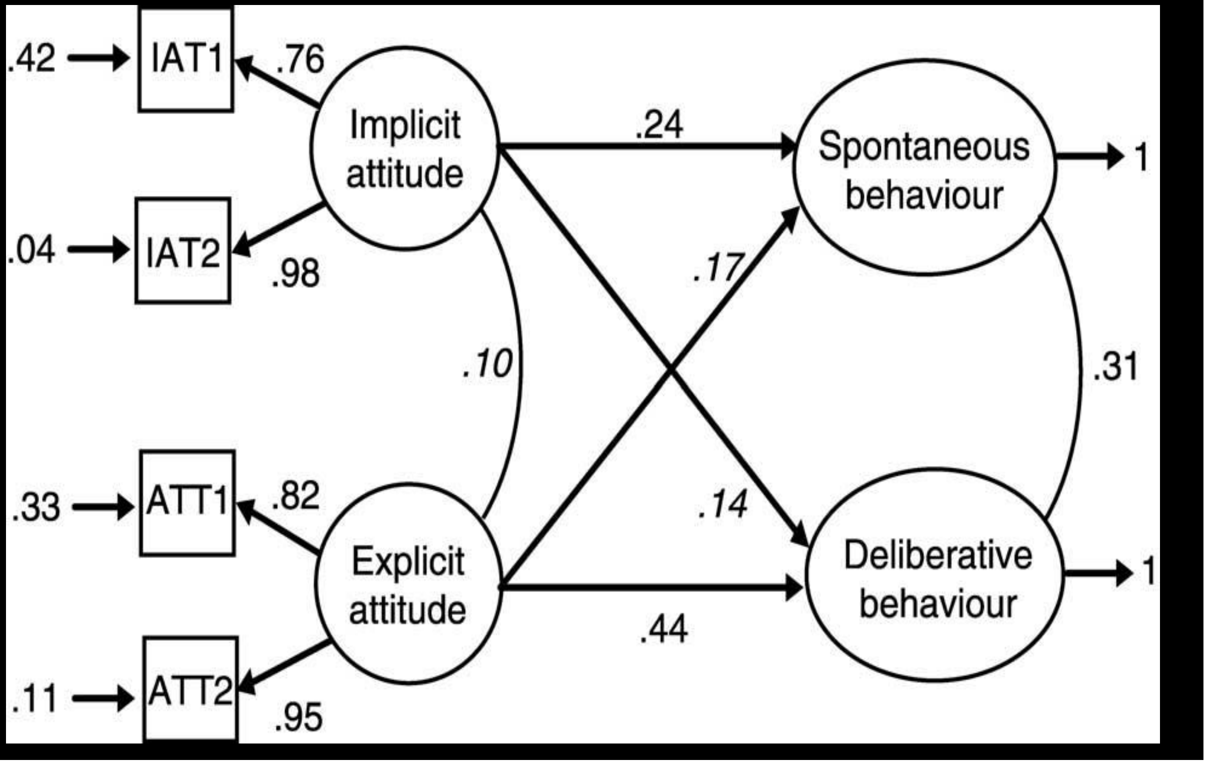
I have argued elsewhere (Rupert 2011*b,* 2015, 2018) that intelligent behavior, including behavior associated with conscious awareness and the first-person perspective, is produced by a system "flattened from above," in which states and processes normally thought to appear at a distinctive (and higher) personal level instead appear alongside (that is, at the same level as) those normally thought to be at the subpersonal level. On this view, the personal level disappears. The entire collection of states and processes associated with the personal and subpersonal contribute, in varying proportions, to the production of human behavior: beliefs, desires, and conscious perceptual experiences sit alongside processes that, for instance, subconsciously assign syntactic structure to the incoming speech stream or analyze variations in light intensity in one's visual field to detect the shape of surfaces in the immediate environment. Accordingly, any of those states might appear side-by-side with any others in the same model of a given set of data.

The point about co-contributing in varying proportions is absolutely essential. One might agree that the so-called personal-level states and processes do not appear at a distinct metaphysical level from sentence-parsers or portions of the visual system that implement edge-detection algorithms, and one might thus accept the letter of the flattened view. But, an objector might nevertheless hold that conscious states or belief-desire pairs, as they appear in the flattened architecture, constitute a relatively isolated domain and make distinctive causal contributions to behavior (*cf.* Drayson 2012, 2014 on the doxastic-subdoxastic construal of the personal-subpersonal distinction) – to such an extent that they can be investigated in essentially the same autonomous way that believers in the personal *level* had thought (see the remarks on folk psychology in Giere 2006, 717). And one might think that, therefore, epistemically speaking, it’s philosophical business as usual.

Note, however, a lacuna in this argument for business as usual. Even if a relatively isolated portion (from a causal-explanatory standpoint) of the flattened cognitive system provides a home to conscious states or to the states adverted to in rationalizing explanations, that fact alone does not underwrite *a prioristic* or armchair epistemic access to said portion of the cognitive system. Now, by my lights, the antecedent is false: the causal-explanatory contributions of states normally associated with the personal level (which we are now agreeing, for the sake of argument at least, does not exist) cannot be very productively isolated from the contributions of states normally associated with the subpersonal level. But, even if my lights shine in the wrong direction in this regard, a case must still be made that *a priori* reflection, introspection, or common sense provides epistemic access to the distinctive portion of the flattened architecture that lines up with the states recognized by folk psychology. That one cluster of factors in a model accounts for nearly all of the variance in a particular set of data, with other aspects of the model sitting idly by, does nothing to establish that humans have introspective or *a priori* access to the former elements of the model, even if it is a model of human behavior.

Generally speaking, my arguments for co-contribution, and the attendant elimination of the distinctively personal level, appeal to methodological naturalism together with prevailing trends in cognitive modeling. Take the data that cognitive science is out to explain, such as differences in reaction times or error rates. Cognitive scientists account for such data by the construction of models. Since a picture is worth a thousand words (or more), consider the following figure, taken from a paper by Marco Perugini (2005).

Figure 2. Standardized parameters for the structural equation model testing for the double dissociation pattern (Study 2 n = 109) (from Marco Perugini, “Predictive models of implicit and explicit attitudes,” *British Journal of Social Psychology*, 2005)



If the reader takes away only one thing from this image, let it be the following. There is nothing in the figure to suggest a difference in levels between the domain of explicit attitudes – those correlated with states supposedly at the personal level – and the domain of implicit attitudes – those commonly associated by philosophers with the subpersonal level. Both kinds of state are part of a network of causes and effects that has no apparent ontological layering built into it. Even if some philosophers would like to insert ontological layers into the picture, the figure itself doesn’t contain them and an assumption of ontological layers does no work in the modeling depicted by the figure, that is, it's nowhere to be found in the statistical modeling of the relations between the various forms of behavior and contributors to them.

It may be worth filling in some details. In the article in question, Perugini reports the results of two experiments, both of which contribute to the now vast body of research on implicit attitudes. The topics of interest in the experiments are attitudes toward smoking and attitudes toward junk food, the basic question being whether subjects' explicitly avowed attitudes toward smoking or junk food govern the relevant forms of behavior or whether, instead, subjects' implicit attitudes do so, and he uses four measures distinct from the experimentally gathered responses to probe independently the presence of implicit and explicit attitudes (IAT1, IAT2, ATT1, and ATT2). In the experimental trials, Perugini explores both deliberative behavior and more time-pressured, spontaneous responses. He models the resulting data using three different approaches – additive, double-dissociative, and interactive. An additive approach takes each of the kinds of attitudes, implicit and explicit, to make a linearly scaling contribution to the production of each kind of behavior (deliberative or spontaneous). A double-dissociative model takes explicit attitudes to produce deliberative behavior and implicit attitudes to produce spontaneous behavior. And, the multiplicative approach takes the two kinds of attitudes to interact (in the statistical sense) in a non-linear way to produce either form of behavior.[4]

The figure included depicts the results of the analysis of the snacks-and-junk-food experiment, which is fit best by a double-dissociative model, the only of the three approaches amenable to an interpretation that reinforces the drawing of a personal-subpersonal distinction.

[4] This nature of the multiplicative model is of central importance, *vis-à-vis* the question of a potentially isolated treatment of folk psychological or conscious states. It's not only that the various elements in the architecture "flattened from above" contribute in varying proportions but that they often do so in such a way that some states modulate the contributions of others. Thus, it is very unlikely that there is a well-circumscribed portion of the flattened architecture in which the previously-thought-to-be-personal-level states appear and out of which a flourishing science can be constructed.

This illustrates the point that, in some cases, the modeling of experimental results jibes with a distinction between levels, even though the model itself is neutral on the matter.[5] The smoking-related data is, however, best fit by a multiplicative model; when trying to predict whether a subject smokes, how much difference is made by a difference in the subject's implicit attitude toward smoking depends on the strength and valence of the subject's explicit attitude toward smoking. For technical reasons, Perugini applies a different form of statistical analysis to his smoking-related data – that is, different from the structural-equation modeling depicted in the figure above in the case of the snacks-related data. But, a distinction between levels plays no role in that statistical analysis, and Perugini makes no effort to layer his presentation of that data, separating streams into levels. So, the flat architecture one sees in the reproduced figure can fairly be taken to indicate Perugini's general conception of the architecture as it pertains to the roles of implicit and explicit attitudes. The contributions of the two streams of processing – whether additive, multiplicative, or even sometimes dissociative – does nothing to support the idea of two different levels in the cognitive system.[6]

In sum, there's nothing in this way of modeling the data that places explicit attitudes at a different level (the personal level), and no other evidence that such a level is playing a causal-explanatory role in such models of human behavior. Instead, there are explicit attitudes and

[5] The fact that there are some cases of the dissociative sort alone does not entail (or highly probabilify) distinct levels of reality. The effects of auditory stimulus can sometimes be dissociated from the effects of visual stimulus. That does not establish that audition and vision are at two different levels of reality! Compare Ned Block's recent discussion of the personal and subpersonal (Block 2017), in which he remarks that there are "clear cases of sub-personal representations (such as gastrointestinal representations) and personal representations (e.g., conscious perceptions)" (*ibid.* 8), and he also expresses sympathy with the idea of a personal *level* (*ibid.*). Although it is not clear that Block himself infers the latter from the former, I contend that such an inference would be misguided. At least across a certain range of contexts, the representations responsible for the production of verbal output might be largely distinct from those that cause quick reactions to oncoming automobile traffic (although one might reasonably doubt the existence of gastrointestinal representations). But, we should not be at all tempted to endorse, on such grounds, a distinction between personal and subpersonal *levels*.

[6] See Gawronski and Bodenhausen (2014, for example at 188–189) for a different kind of approach to implicit attitudes research that also supports the skeptical conclusion reached here concerning the relation between implicit and explicit attitudes, on the one hand, and the personal-subpersonal levels, on the other.

implicit attitudes, each of which can contribute to the production of either kind of behavior: deliberative or spontaneous. And, each of the two kinds of attitudes can, and sometimes do, modulate the contribution of the other with regard to either kind of behavior. This last point is of special relevance, for it suggests that, even if we limit ourselves to thinking of the personal level as a level of description, a self-standing vocabulary of the sort Dennett has in mind, contemporary cognitive science cuts against the idea that the vocabulary in question fruitfully segments a domain of explanation that does distinctive work in isolation, or enough distinctive work to elevate it to a privileged status in the way proponents of a personal level seem to have in mind (McDowell 1994).

Although space does not allow consideration of the full range of objections that might be made, I consider one. This may give the reader a deeper appreciation of the flattened view and of how it might be defended against objections.

It is sometimes claimed that the role of cognitive science is to deliver a so-called vertical explanation (Hornsby 2000, Bermudez 2000, Drayson 2012) of personal-level capacities (or, even weaker, an enabling explanation of such capacities or states – McDowell 1994). A vertical explanation is an account of an existing capacity in terms of subcapacities (often assumed to be at a lower level). This is to be contrasted with so-called horizontal explanation, which accounts for a token concrete event – for instance, a token action – by appeal to token causal antecedents of that event. On this view of cognitive science, the enterprise presupposes a personal level in its target *explananda*, the character of which are delivered *a priori* or by common sense, and the job of cognitive science is *not* to give horizontal explanation of token events – personal-level explanations do that – but only to give vertical explanations of personal-level abilities or capacities.

As philosophy of science, this picture seems fundamentally misguided. Cognitive science is a science, responsive to data. As Newell, Shaw, and Simon put it, at the dawn of cognitive science, “What questions should a theory of problem solving answer? First, it should predict the performance of a problem solver handling specified tasks” (1958, 151). For a contemporary discussion that is unremarkable in the way it assumes the same view, as a matter of course, see Botvinick and Cohen (2014, for instance at p. 1255). First and foremost, cognitive science offers horizontal (causal) explanations of measurable behavioral data. Relations between hidden quantities – as elements in a model – (causally) explain instances of human behavior (the data to be modeled); they do so on the assumption that the values of those quantities correspond to token “hidden” states, processes, and their properties that produce the behavioral tokens in question, in the standard horizontal way (De Houwer, Gawronski, and Barnes-Holmes 2013).

Of course, data do not wear their best model on their face. Thus, it is possible that the best model of a given set of data in cognitive science – or more likely a collection of sets of data – has, built into it, structures that reify, or are best interpreted as reifying, personal-level capacities. But, that is to understand the cognitive-scientific enterprise in a way that turns the objector’s picture on its head. In that case, personal-level capacities would not be *explananda*; rather, they would enter the scene only as part of the *explanans*; they would have been introduced – explicitly or implicitly – into models of the data to do causal-explanatory work, in just the way any theoretical entity, force, or relation might be introduced. That being said, so far as I can tell, they aren’t even *explanans* in the current state of cognitive science.

This is not to deny that general talk of human capacities sometimes plays a role in the process of generating new ideas for experimental designs or as a convenient way of summarizing patterns in models or results. What has model-disconfirming or model-rejecting power, however, are data on performance, not independently arrived at claims about the personal-level facts. I find no evidence that it is part of the methodology of working cognitive science that the cognitive scientific community chooses an otherwise worse model of the data over a better model of the data because the former does a better job of explaining an "independently given" fact about the personal level, something about a personal-level capacity, for example. In this sense – that is, in the sense in which *explananda* must play an "adequacy-condition-setting" role in model selection – personal-level facts are not the *explananda* of cognitive science.

Notice that the flattened view is not inherently eliminativist about folk psychology (Churchland 1981). Revisit Perugini's diagram above. Explicit attitudes appear in that diagram; they are not eliminated. Of course, cognitive science could eventually eliminate folk psychological states, e.g., by incrementally changing its models of the processes that produce intelligent behavior, reaching a point at which no components in those processes have many of the properties thought by the folk to attach to beliefs, desires, etc. (Stich 1996). That remains to be seen (*cf.* Burge 2010, where the idea that cognitive science will do without propositional attitudes is dismissed, with no argument, as "outlandish" – 40). Regarding folk states, the present point is conditional: if folk attitudes are vindicated, it will mostly likely be as non-isolated contributors to the production of intelligent behavior, at the same level as so-called subpersonal states or processes, not as items in a distinctive domain – the personal level – about which philosophers can amass self-knowledge by deploying only introspection, common sense, or *a priori* reflection.

IV. The Object of Self-Knowledge in a Cognitive System Flattened from Above

Let us accept, then, that the mind is flattened from above. What becomes of the quest for self-knowledge? In this section and the one to follow, I develop a framework within which to seek self-knowledge and to understand the enterprise of seeking it.[7]

Self-knowledge involves both the self and knowledge. Let us ask here about the self. What is the self about which one might have knowledge? Consider two possibilities. According to the first, the self is the cognitive system as a whole. What is the cognitive system as a whole? It is the relatively integrated, relatively persisting collection of mechanisms that contribute in overlapping subsets to the production of a wide range of forms of intelligent behavior. In previous work (Rupert 2009, 2010, 2019), I've characterized integrated systems using a statistical measure of the clustering of contributing mechanisms, according to how likely a given mechanism is to contribute to various forms of intelligent behavior conditional on the contribution of other mechanisms. But, in the present context, discussion of such details would take us too far afield. In place of a detailed specification, our working notion is that of a cognitive architecture (Pylyshyn 1984), construed broadly, for the architecture might be computationalist, connectionist, dynamicist, brute neural (or biological), subsumption-based, or whatever (Rupert 2009).

Roughly speaking, such a system encompasses everything that contributes, in a consistent, distinctive, and flexible way, to the production of intelligent behavior. And although

[7] This framework in question is motivated by and dovetails nicely with the view that the mind is flattened from above; but neither view (the framework in question or the view that the mind is flattened from above) alone entails the other. Because the framework doesn't alone entail the flattened view, the defender of the personal level might, without inconsistency, claim that her view, properly adjusted, can accommodate what I say below about self-knowledge and self-improvement. As I see things, such accommodation would come at too great a cost, a forfeiting of what was meant to be distinctive and philosophically useful in claims to the existence of a personal level. Moreover, given the powerful, independent arguments against the existence of a personal level, it strikes me as quixotic to develop a way of understanding the personal level that renders a commitment to the personal level consistent with the views presented here about self-knowledge and self-improvement.

Here's a more upbeat (and dialectically neutral) way to think about the takeaway message of this footnote. One needn't have made up one's mind about the existence of a robust personal level in order to find provocative or promising what's said in the remainder.

this sketch stands in need of elucidation, a fleshed-out conception of cognitive architecture will surely include all sorts of mechanisms and processes that don’t seem particularly self-related. For example, processes governing the relative timing of the contributions of syntactic and semantic analyses to sentence processing (Kim and Sikos 2011) may constitute important workings of the cognitive system, but they don’t seem *self*-related; it seems strained to call knowledge of such processes ‘self-knowledge’, in contrast to such clearly self-related pieces of knowledge, for instance, as knowledge of one’s preference for certain pastimes or of one’s emotional dispositions in the context of interpersonal relationships.

Consider, then, a second possibility, according to which the self is proper subset of the collection of mechanisms and processes included in the entire cognitive system, in particular, the distinctively self-related ones. The strategy here is to identify a set of mechanisms or structures within the cognitive system that play the distinctive role of the self, in, for instance, the

-fixing and updating of doxastic-like commitments

-maintenance of a coherent biography, a narrative about one’s self and one’s life

-explaining or defending of one’s actions to others

-issuing of reports on one’s own success or failure, on short and long time-scales (e.g., the issuing of metacognitive judgments)

-production of inner speech

-initiation and guidance of planning and decision-making

-initiation and guidance of attentive motor control and skill-execution

This second possibility might seem to be the obvious frontrunner. But, I’m skeptical. The cognitive system is a complex collection of mechanisms and processes, densely and richly interconnected and co-contributing in shifting subsets to the production of self-related behavior. Given the nature of the system, there's unlikely to be any principled distinction between the parts that do the "selfy stuff" and the parts that do not.[8] In recommending the second possibility, it's as if someone were telling us to locate the part of the Internet that is the political part. Some sort of sophisticated network analysis could almost certainly identify clusters of Web sites that present particularly political content, but these clusterings would be far from tidy, and the analysis of such correlations would almost certainly reveal vast amounts of off-topic content, partial interaction between the more political sites and the less political sites, and so on. The status of distinctively self-y parts – as a proper subportion of the cognitive system – would, in fact, seem to be like that.

Real human cognition is messy, filled with cross-talk, redundancy, and competitive processing, and affected by fine-grained details of embodiment, among other complicating factors. And, although significant distinctions may emerge from cognitive-scientific modeling (Wilson 2002, Rupert 2009), we should remain open to the construction of hybrids models – models that include both connectionist and computationalist elements, models that include both implicit and explicit attitudes, models that include both slow serial processes and fast parallel processes, models that contain terms both for interference from semantic association during reasoning and for the natural-deduction rule being applied during that same step in reasoning.[9]

[8] I am not claiming that it's a messy and indeterminate matter whether a personal level exists or where the boundary lies between the personal and subpersonal levels. *That*, I have argued, is a determinate, non-messy matter; no personal level exists, so far as I can tell, given the current state of the evidence. But, once commitment to the personal level has been abandoned, the question arises whether there is a well-defined self that exists as a proper part of what happens or appears at the so-called subpersonal level. I'm suggesting that the answer to this question is, "No, in the flattened picture, there is, at best, a messy, indeterminate distinction between, on the one hand, the self-y parts of the processes and states that appear at the single level in question and, on the other hand, and the non-self-y parts appearing at that same level."

[9] In other work, I've proposed a statistical method for distinguishing the proper cognitive system from other contributors to the production of intelligent behavior (Rupert 2009, 2010, 2019), a measure that clusters together the mechanisms that distinctively contribute to the production of intelligent behavior and excludes one-offs and special-

Defenders of a distinctively self-related subportion of the cognitive system might appeal to the narrative conception of the self (Dennett 1991, Schechtman 2007, 2011). Perhaps what is distinctively self-y are the stories we tell about ourselves, to ourselves and to others. In which case, perhaps the self-y parts of the cognitive system should be identified specifically with the mechanisms or module(s) responsible for encoding and generating said stories. In which case, the self just is the well-delineated proper part of the so-called subpersonal level that encodes – and processes the encodings of – the content of the individual's biography or life story.

But, this won't do. The content of such narratives varies from telling to telling (Dennett 1991), with a variety of contextual factors determining what version of that specific model gets activated. Consider the extensive research on the reconstructive nature of memory -- Shacter 2012 – which applies here, given that some significant portion of any of the self-models will itself be encoded in memory.

And, it's not just a matter of one reconstructing one's story differently on different occasions for different audiences – that is, not just a matter of having a single narrative from which is derived different output on different occasions. That much is true, so far as it goes. But, to complicate matters further, a single cognitive system is likely to contain *different* self-representations – or self-y parts of the cognitive system, or self-models (Flanagan 1994) – encoded in different parts of the brain, and playing different self-related computational, behavioral, or otherwise processing-related roles in the cognitive system (Bickle 2003). This makes it difficult to pin down *the* narrative that constitutes the self. And, to make matters even

purpose tools. Might a version of that proposal apply here, allowing us to carve out a distinct subportion of the cognitive system that uniquely plays the role of the self? I'm skeptical, for reasons given in the main text. If one were to pool together all of the behavior that is potentially self-related (all of the behavior connected to the bulleted list above), and then apply the statistical measure in question, in hopes of identifying the cluster of mechanisms that contribute distinctively to the processes producing that behavior, I suspect the output would be the set of mechanisms that constitute the entire cognitive system, or something so near to it that the difference would not alter the thrust of the arguments given or the gist of the picture presented in the remainder.

less tidy, each of these many self-models is likely to (a) share ancillary resources, to do with attention and so on, with processes deploying other self-models and (b) to be influenced in the content it produces on a given occasion by factors what is normally thought of as subpersonal processing – from ambient smells, to available attention, to pressing goals, to hormone-controlled release of neuromodulators.[10] Thus, the search for a single, privileged self-y part of the cognitive system is triply hampered by forms of context-sensitivity, fragmentation, and variation: a single cognitive system contains multiple self-models, different ones contributing to different self-related tasks; each individual self-model contributes differently (to the extent of differing in its content) on different occasions; and what might reasonably be treated as distinct self-models contribute to the performance of self-related roles in tandem with various other resources, shared to various degrees across various contexts with the causal processes by which other self-models perform their self-related roles.[11] For all of these reasons, it seems artificial to try to delineate firmly the properly "self-y" subset of mechanisms responsible for the production of narratives from those mechanisms that are not part of the self.

[10] Note too that the entire cognitive system may change its character even on fairly short time-scales, given, for example, neuromodulator-triggered changes in effective connectivity (Anderson 2014). We might thus wonder whether the only consistent element that might play the role of *the* self, to be known, is not the cognitive system but the physical basis of the disposition to shift from one cognitive system to another (from one complete network of effective connectivity to another) in response to context.

[11] These remarks dovetail with some aspects of the views of Flanagan (1994), Velleman (2005), and Carruthers and Williams (manuscript), all of whom emphasize in their own way the forward-looking contributions of narrative or other forms of self-model in cognitive processing. My point is to emphasize the existence, in any given cognitive system, of many such models (not all of which are linguistically encoded), different of which contribute to distinct aspects of self-related processing – different models, as one might say, filling different self-related computational or causal-functional roles – and do so in significantly context-variable ways. (Velleman takes the first step down this road: "we tell many small, disconnected stories about ourselves – short episodes that do not get incorporated into our life stories" – 2005, pp. 72–73.) It's less clear how my view jibes with Schechtman's (2007, 2011). She emphasizes the ways in which the content of narrative colors the subject's conscious experience, which I don't take as an *explanandum*. It is quite possible, however, that one's conscious experience reflects differences in a self-model – whether a robustly available, linguistically encoded model or something more like a proprioception-based body schema (Gallagher 2005) – which differences affect the production of observable, self-related behavior. In that case, such a view might productively be brought into contact with the one developed here.

Thus, although I don’t think it’s utterly hopeless, I despair of any attempt to specify a determinate boundary between *the* self-constituting component of the cognitive system and those components of the cognitive system on the other side of that boundary. In what follows, then, I treat the entire cognitive system, as well as any of its proper parts, as potential objects of self-knowledge.[12] Of course, I acknowledge that we tend, in self-related contexts, to care more about some aspects of the cognitive system than others. This concession carries only so much weight, though, and seems unlikely to help us to carve out a proper part of the cognitive system as *the* self, for the parts of the cognitive system that we tend to care about, when we’re concerned about the self, vary quite a bit. To the extent, for example, that we’re interested in the ways in which subjects use a self-narrative to justify their potentially morally questionable actions to others, after the fact, we might be especially interested in a module that encodes information about local moral norms. But, to the extent that we want to know why someone is the kind of person who frequently interrupts others during conversations, our interest might be satisfied by information about idiosyncrasies in that person’s linguistic processing mechanisms and the way those mechanisms interact with that person’s model of conversational turn-taking. Given how much variety and flexibility such explanatory contexts exhibit, I continue to doubt that a single, consistent self can be carved out of the cognitive system, in a principled way.

V. Self-knowledge, Sometimes without Belief or Truth

What about knowledge? Generally speaking, I treat the justification- and warrant-related aspects of knowledge in the spirit of reliabilism (Goldman 1979). I do not defend this broadly reliabilist attitude here or say much more about it beyond this: the arguments in support of naturalism are

[12] After all, any property of a part can be conceived of as a property of the whole. To put this point in neural terms, the property of having neurons active in V1 is a property of the brain, viz. its having neurons active in V1. In brute physiological terms, my hand’s having five fingers is a property of my entire organism; this body has the property of having a hand with five fingers.

powerful, and the natural world would seem, at least above the level of fundamental physics (and possibly even there), to be largely a flux of probabilistically related event-types; thus, if justification is a natural kind, property, or relation – genuinely part of the natural order – it seems likely to be built out of probabilistically related event-types, which captures well enough the spirit of reliabilist justification.

In keeping with this view of the universe – as being constituted (primarily) by patterns of probabilistically related event-types – I leave open the possibility that possession of self-knowledge does not require the subject to possess a corresponding belief, partly because I hope to identify a notion of self-knowledge that can be preserved even if there are no beliefs. (I don't presume eliminativism about folk psychological states, but it may well be true.) A related consideration runs as follows. It might be that belief attributions are sometimes true, but are, from the standpoint of a mature cognitive science, made true by psychological states that do not form a natural kind. They might, for example, all be instances of a single but more expansive kind; and it might be that states serving as truth-makers of belief-attributions (when they are true) do not bear any distinctive, scientifically important mark as a subset of that more inclusive kind: from the standpoint of cognitive science, there may be no important difference between, on the one hand, the psychological states that serve as truth-makers for true belief-attributions and, on the other hand, states of the same natural kind but that do not serve as truth-makers of true belief-attributions. And, in that case, the broader kind might be of greater causal-explanatory value than the parochial, nominal kind; even if belief-states (that is, the states that make true belief-attributions true) possess some distinctive properties, such properties might be of minimal interest; the more powerful and interesting scientific kind might be the kind that includes both belief-states and others of the same kind but that do not play the role of truth-makers of true belief-attributions. Perhaps, these other states carry information about the cognitive system without being truth-evaluable. For causal-explanatory purposes – for the purposes of predicting or intervening on behavior, for example – it may be that the information a state bears about the cognitive system (regardless of whether that state is true, or even truth-evaluable) is of greater scientific utility than its status as a state that makes (or doesn't make) an everyday belief-attribution true. From the standpoint of getting people to behave in a certain way, it may be more important what a given state is correlated with statistically than what its content is or whether it has content. I do not argue or presuppose that cognitive science will deliver such results, only that the current state of evidence provides sufficient reason to develop an approach to self-knowledge inclusive enough to apply if this eventuality obtains.

How should we think about non-belief-involving knowledge states? Some cases may involve know-how or what's sometimes called 'procedural knowledge', but I do not defend the exclusive importance of that category as such. Also important from the current standpoint are bodily manifestations of self-knowledge. For example, it might count as self-knowledge if one's hesitancy to attempt a certain task correlates with one's not being in sufficiently good physical condition to perform that task, regardless of whether one can report accurately on why one didn't attempt the task or on one's own physical condition.[13]

[13] Allowing states to count as encodings of self-knowledge merely in virtue of their information-carrying properties allows for a degenerate form of immediate self-knowledge. For instance, every time one speaks, the vocalization carries information that the system is in the cognitive state – whatever it might be – that produced the vocalization. (Compare: A burp might be a perfect indicator of the physiological state that was primarily responsible for the burping, but the content of the "utterance" tells us nothing interesting about said physiological state and what other kinds of behavioral output that state might cause – *cf.* Dennett 1969, 103.) I don't take this to be a shortcoming of the current approach, for a more traditional approach to self-knowledge – say, one that requires that self-knowledge be encoded in a true belief – also allows for cheap, uninteresting self-knowledge. When one utters, "I am in the internal state that is producing this very instance of verbal output," one has thereby made a true, justified, self-related report. One might worry that there is significantly more cheap, uninteresting self-knowledge on the mere-information-carrying view than on a more traditional view. Perhaps, but tricks are available to construct a vast amount of self-knowledge consistent with the traditional approach: "I am in the verbal-output-producing state that is the verbal-output-producing state that immediately follows, in a series of verbal-output-producing states, the one that produced my most recent instance of verbal output," and so on, with nested iterations.

In other unorthodox cases, beliefs are indeed involved, but they are false, or at least deeply uninformative *vis-à-vis* the states of which the subject thereby has knowledge. In these cases, the occurrent activation or avowal of the belief correlates highly with a particular state or process of the cognitive system, irrespective of the content of the belief in question. In such cases, self-knowledge is encoded in a truth-evaluable vehicle, but possessing self-knowledge doesn't depend on the content of that vehicle; having content is an accidental or tangential feature of the knowledge-vehicle, and thus so is its truth (or falsity). For example, after experimental subjects have learned new material, they gradually undergo a well-documented Remember-to-Know shift. In a typical experimental design (Dewhurst, Conway, and Brandt 2009), subjects study unfamiliar content and then, at intervals following study, are asked questions about the content itself. They are also asked metacognitive questions about the answers they gave to the first-order questions about content: "Did you 'remember' the answer or did you 'just know' it?" Subjects more commonly report that they "just know" their correctly given answers at later testing times than they did at earlier testing times, and are more likely to report that they "remembered" their correct answers at earlier testing times. This is best explained by the association of the alternative "Remember" with detailed episodic memory, in contrast to "Know," which correlates with the shift in the representation of the content in question to semantic memory (that is, its being re-encoded in the abstract form in which humans store generalized, factual information about, say, relations between types or categories). Thus, their metajudgment "Just Know" indicates the completion of a certain cognitive process or the presence of a new state with certain relations to other states (deductive relations, perhaps, but not relations to states with concrete-imagistic-time-stamped properties). So, in general, then, subjects' "Just Know" reports reflect something important about their cognitive system, but not something that subjects themselves can report much about (depending perhaps on whether they've completed a course in cognitive psychology!). And, if eliminativism about folk psychological states is true – and there are no beliefs – then humans have no knowledge at all as traditionally construed (as justified true belief); in which case, the subjects' reports in these cases – that is "I just know" – would seem to be straightforwardly false.[14] But, from a pure information-carrying standpoint (Dretske 1981), the subjects "know" something about their own operation: differences in their metacognitive reports correlate reliably with differences in the processes that produce the first-order task responses at issue.[15]

To be clear, though, nothing I've said in this section is meant to disallow cases in which self-knowledge is encoded in justified true beliefs. In other words, I propose that we allow a variety of kinds of vehicles to express or encode self-knowledge and that we admit both true-description and state-tracking (or information-carrying) as legitimate accuracy-related components of the (now liberalized) JTB formula.

VI. Coordination and Self-improvement

Humans value self-knowledge for a variety of reasons. Many take it to be intrinsically valuable; others want to know their own capacities and limitations so as to facilitate the setting of realistic goals for themselves; and so on. In this section, I consider only one such value, the value placed

[14] Whether "just know" reports would be false under such conditions depends largely on contentious semantic issues and possibly the everyday meaning of 'to know'. If (a) content is to a significant extent descriptivist or depends on inferential role, (b) the folk concept of knowledge is of a state that entails belief and (c) there are no beliefs, then all knowledge, at least in the mouths of the subjects of the experiments in question, are likely false. If, however, the content-determining aspect of content is referential and reference is, say, causally determined, and causal relations hold only between natural kinds, it's possible that a knowledge attribution in the mouths of the subjects be true, even if there are no beliefs; knowledge may well be a natural kind, instances of which are causally responsible – in the reference-fixing way – for the everyday use of 'know'. But again, I would insist that these matters need not be sorted out. It's not the truth of a subject's report that's relevant. What's relevant is that the subject's metajudgment "Just Know" tracks a state of her cognitive system that is itself correlated with a variety of other behavioral and physiological measures.

[15] It might be well advised to employ a different term, 'self-indication', perhaps, following Dretske's use of 'indication' (Dretske 1988), leaving 'self-knowledge' to those with more traditional views about the psychological state involved. Yet, in my view, naturalistically leaning epistemologists should consider a broad range of states as potential vehicles of knowledge, properly speaking, and so hesitate to concede the use of 'self-knowledge'.

on self-improvement. If the human psyche is flattened from above and if the self and self-knowledge take the forms proposed in the two preceding sections, how might self-knowledge lead to self-improvement? I propose that it involves distinctive forms of coordination among, and interaction between, states and processes commonly associated with the self, on the one hand, and states and processes typically associated with the subpersonal level, on the other.

Although it involves gross oversimplification, it will be useful to set the discussion within the framework of a dual-systems approach to cognitive processing (Evans and Frankish 2009).[16] The dual-systems literature distinguishes between processing that is fast, automatic, associative, relatively effortless, changes gradually with habituation, and that places little demand on working memory (System One style processing, as the terminology goes) and processing that is slow, deliberate, serial, conscious, effortful, allows for one-shot learning, and that makes heavy demands on working memory (System Two style processing).

Using the language of these two processing styles, one important way in which self-knowledge can guide self-improvement is by System Two's acting as coach and trainer to System One: *States in System Two can come to know (or be the knowledge states that represent) what is happening in System One either by encoding merely correlational information (regardless of whether the states in question have content) or by encoding accurate content, and System Two's doing so provides a salutary nudges to System-One processing, for instance, by engaging a filter (in the mathematical sense) that sharpens signals internal to System One processes, thereby improving the efficiency or effectiveness of the relevant System-One processing.*

Some comments are in order. The processes of self-improvement I have in mind exploit both the minimal representational resources in play in System Two processing as well as the

[16] Some of the essays in Evans and Frankish (2009) discuss shortcomings of a strict dual-systems approaches.

more sophisticated, propositional representations also appearing in System Two.[17] Consider an example of the former case, in which self-improvement results from System Two's ability merely to carry information about what's happening in System One, regardless of whether the System Two structures in question encode propositions about occurrences in System One or whether those encodings, if they do have truth-evaluable content, are true. System Two might help to improve musical performance by producing encouraging verbalizations in response to especially well-constructed musical phrases ("That's right!" "Do it!" or "Uh-huh"). But in order to do so effectively, that is, in a way that improves the playing of the instrument, the differences between the cases in which System Two produces "Uh huh" and the cases in which it produces contrasting output (e.g., "Doh!") must track – though not necessarily perfectly – differences between System One processes that produce especially well-constructed musical phrases and System One processes that do not. Notice, too, that such signals can be nested. Variation in a higher-order signal in System Two might carry the information that System Two is now in a state of a sort such that, when in the recent past a state of that sort issued in a training signal, that signal was highly effective (measured by performance on immediately subsequent performances); in which case, System Two will amplify the training signal it sends to System One in the present case. This could be thought of as a metacognitive signal of a sort, a level of confidence in System Two's inclination to send "Uh-huh," in situations such as the present one,

[17] Though arguably not appearing exclusively in System Two; for, much of what has traditionally been thought of as subpersonal processing involves the manipulation of propositional representations. Consider the fast, automatic way in which a skilled reader processes and comprehends text (Gathercole and Baddeley 1993, Ch. 8); the maintenance and updating of a coherent narrative or text model is difficult to make sense of without invoking propositional structure. Similarly, dissonance reduction – roughly, the adjusting of one's conception of oneself or of one's own cognitive processing in response to behavior the natural explanation of which would not be flattering to oneself or does not fit one's conception of oneself – is normally thought to be a subpersonal process, but it is difficult to explain in purely associationist (that is, nonpropositional) terms (Quilty-Dunn 2020). See also Mandelbaum (2015) for arguments that propositional structure is at work in forms implicit cognitive processing associated with implicit bias.

as feedback to System One, which thereby manifests itself in more robust reinforcement of System One's processing than it would have otherwise.

Aspects of the preceding example might shed light on another main aspect of the view, the idea of sharpening, as a mechanism by which, in virtue of its possession of self-knowledge, System Two has a salutary effect on System One processing. I gave the example above of the application of a filter, which might, for example, smooth out a motor control signal or sharpen boundaries between segments of a multi-step or multi-phase task in System One, as in the serial, distinct activation of digits in piano-playing. But, I don't intend to limit the range of possibilities in this regard. Feedback signals can play a role in learning in cases in which the adjustments made in response to feedback are not best modeled as the effects of the application of a filter. The more generic notion is that of reinforcement learning.

On this picture, activity in System Two helps to make System One's processing more efficient, partly by way of System Two's real-time guidance but also by its giving after-the-fact "tutelage." Such tutelage might result in changes of a structural sort; a higher-order signal in System Two might help to effect better coordination between System One and System Two, by, for instance, affecting parameter settings in the channels of communication between the two systems.[18] After-the-fact tutelage might amount to the conscious replaying of a representation of what happened in an instance in which System One produced a failed action. Or, in a case of successful action, it provides the opportunity for language-based rehearsal and thus reinforcement of whatever System-One process was carried out. It can be a way of amplifying a training or error-detection signal operative in System One. And, the efficacy of this process can sometimes turn on the accurate representation in System Two – whether arrived at in some

[18] And note that System Two processes or structures can do so regardless of whether the content of the System Two representations in question constitutes a narrative; in other words, the points here are orthogonal to the central aspects of the debate between Schechtman and Strawson (Schechtman 2011, pp. 407–410).

immediate causal channel or by investigation from a third-person perspective – of features of System One. Sometimes self-directed talk helps to train up System One partly because the self-directed talk expresses independently arrived at truths about System One.

Perhaps a certain human ideal appears in this vicinity, realized when an organism achieves widespread coordination of System One activities and System Two's representations of them. Consider a subject who exists at the end of cognitive-scientific enquiry and who has mastered all there is to know about cognitive science; she moves through her life running through a series of System Two states that explicitly represent and accurately describe the System One processes she simultaneously executes, exhibiting a kind of perfectly naturalized self-knowledge. Such System Two representations of the activities System One would be justified by the doing of good science, the doing of which itself requires coordination between System One activities (experimental design, execution, and interpretation) and System Two activities (explicit aspects of scientific reasoning). In this case, the coordinated collaboration of System Two and System One effects further coordination between System Two and System One. This might come to pass not only because an enlightened subject has learned everything there is to know about her own System One; it may also be a top-down effect of one's coming to believe certain theories about System One that System One then is more likely to operate in a way that satisfies those beliefs – entrained, as it were, by the System Two structures with which it interacts. To co-opt Gendler's terms, and to subvert them a bit, perhaps if one knows enough cognitive science, one can rid oneself of alief-discordant beliefs, both by acquiring true beliefs about System One and by acquiring beliefs about System One that help to make it the case, by causal influence of System One, that those beliefs become true.

I should emphasize how widespread the activities of System One are, beyond the familiar domains of athletic skill and musical performance. Consider, too, oratory, the giving of lectures or presentations, typing, participating in a conversation, making jokes or witty remarks in real time, running meetings, in the carrying-out of a wide range of executive-function tasks that involve prioritization or the allocation of attention, and more. Most (perhaps all) human cognition has a significant component of System-One-based skill to it, for most of us, most of the time, performed with virtuosity. System One is mostly amazingly good at what it does. In light of this, one might wonder, then, whether System One has anything to teach System Two.

In this regard, consider Fiery Cushman's (2020) view of rationalization, according to which rationalization drives what he calls "representational exchange" (2020, 3). According to Cushman, our behavior is often produced, to at least some significant extent, by habit, instinct, and other non-rational processes – that is, processes typically associated with the subpersonal and with so-called System One. Frequently, after such a process has produced behavior, the subject engages in rationalization, recasting the behavior-producing processes as rational, as having been driven by beliefs and desires that did not, in fact, contribute causally. This much is relatively uncontroversial. Cushman adds the claim that rationalization can have salutary effects, by providing the mind with useful new beliefs or desires. Assume that the nonrational forces at work are adaptive or useful, perhaps, in the case of instincts, because they were selected for evolutionarily (in an environment near enough to our current one). Then, even though the nonrational process in question did not include the beliefs or desires in question, a different way to get the same benefit from the relevant action would be to possess and act upon those beliefs and desires. Rationalization of behavior – the subject's concocting a set of beliefs and desires that would have led to the behavior in question – can be a way for her to discover new beliefs and desires the adoption of which allows her to reap, more flexibly, the same benefits provided by the instinct-driven behavior that's being rationalized (and that was not driven by the beliefs and desires in question).

Cushman asks his readers to consider an analogy. An infant pauses at the edge of a visual cliff and reverses direction. This might be a matter of instinct, an evolutionary adaptation that protected the infants' ancestors from fatal falls. But, if the infant were capable of rationalization, she might attribute to herself the belief that cliffs are dangerous. Her actual adoption of that belief, for future use, would be highly adaptive, and could help her to avoid cliffs even in cases in which the cliff's edge is not staring her in the face. Of course, the infant doesn't engage in rationalization. But, for those of us who are in a position to rationalize our instinct or habit driven behavior, our doing so can, Cushman claims, be adaptive, by providing the sort of benefit illustrated by the case of the infant on the visual cliff.

On one way of understanding Cushman's view, it involves System One's tutoring System Two (*cf*. Cushman 2020, 52). On my view, this perspective should be expanded to include all sorts of cases in which System One style processing produces outputs – for example, makes reliable discriminations – that can, in turn, be encoded explicitly, in widely accessible representations that can thereafter be deployed flexibly and (in some cases) control verbal report. System One acts on the world in ways not driven by explicitly encoded beliefs and desires (that is, by information-bearing or motivational states that can be widely deployed and that can control verbal report, including verbal report of experienced inner speech during slow, careful reasoning). Sometimes System Two learns about the way the world is or works or what possible patterns it contains by attending to the products of System One's activities. In exercising skills – from sculpting to shooting baskets to speaking in public – System One produces results that exhibit various properties, including patterns in those properties. Shots go in from the foul line with feet set firmly on the ground, but less often when the shot is attempted while on the run. System Two can come to appreciate that the distinction between a static body and a body in off-balanced motion is an important distinction to consider in planning a strategy for a game or devising with new plays to run. This can give System Two all manner of new concepts or conscious contents, beyond what might be acquired as a side-effect of rationalization, à la Cushman's view.

Consider the extent to which feedback loops can occur. Imagine an example from mathematics. Having attended to the processes and outcomes at work in the solving of a series of math problems, System Two explicitly encodes heuristics that can be used to approach new problems (to tell whether this problem is best handled by factoring into an explicitly represented "shared term", now place in front of parentheses, or whether instead by fractional representation and cross-cancellation). In this case, System One generates various problem-solving strategies in an intuitive or automatic way, the patterns in which System Two sees only after the fact, and which System Two then encodes explicitly for use (cf. Karmiloff-Smith on representational redescription – Karmiloff-Smith 1992, Clark and Karmiloff-Smith 1993). System Two can then explicitly guide the solving of future problems, by, say, attending to the perceived differences between the cases in which one kind of factoring process worked more effectively than another. System One may then, when following System Two's guidance, exhibit a more refined pattern of differences in its performance – perhaps on new, related kinds of problems – which System Two might then notice and repackage in explicit format, and so on. In what way is System One tracking states of System Two, thereby acquiring knowledge of System Two? By being systematically responsive to differences in states of System Two, in some cases, in ways that reflect differences in the content of those states of System Two. (And, notice a further way in which System Two might harness knowledge about System One; it might be more likely to learn from System One's outputs if it represents them explicitly as its own outputs, that is, as products of the activity of the very organism of which it is itself a part.)

Many readers may recoil from this view, for, in the contemporary context, System One is often seen as a moral culprit, the domain of the implicit attitudes that provides the last and most resistant bastion for racism, sexism, and classism. Consider Tamar Gendler's (2008*b*) discussion of the concordance of aliefs and beliefs, including in morally charged contexts, in particular, in connection with implicit racial prejudice. A strong presumption runs through her view, and it's often thought to be of ancient provenance: that self-improvement requires us to bring our aliefs (essentially System One states and processes) into concordance with our reflectively held beliefs (states in System Two). On this view, the relation between System One and System Two is the relation between, on the one hand, reason – expressing our most considered picture of who we are and our most noble and enlightened image of who we want to be – and, on the other hand, the passions – ill-advised and dangerous, and that must be tamed and kept in check. On this view, as applied to the case of implicit racial bias, we hope that System Two can rise to the challenge of "disciplining" racist aliefs so as to bring them into line with (or at least keep them out of the way of) our well-considered beliefs about morality and social justice.

I resist this picture, as a general picture of the relation between System One and System Two. In many cases, self-knowledge helps to bring System Two into alignment with System One, the converse of relation Gendler envisions. Socially sensitive System One style processing can deliver a perception that one situation involves something disturbing that another situation does not, which can be enormously useful in training System Two to encode explicitly the morally relevant aspects of situations. In such cases, it is not conscious, deliberative reason that needs to enforce its rationally arrived at conclusions on the passions of System One (Gendler 2008b, 578). It is more a matter of reason's learning to trust and support the passions (and the discriminations made by System One), of the coordinated system as a whole learning to trust, for example, System One's repulsion at seeing George Floyd killed on video, and to then formulate and apply a label or concept accordingly (if one is not already available). Moral reasoning itself is shot through with the contributions of System One. It is only (or at least largely) because System One has in the past done well in the moral domain that we recognize that something has gone wrong in cases in which deference to System One seems to lead us astray – for instance, when implicit bias rears its head. It is because System One contributes so effectively in the context of moral and social cognition that we can see that something's going wrong in pathological cases involving, for example, System One's contribution to acts of racial discrimination. System Two rings its hands about System One's bad behavior only because System One, when doing its successful work, has trained up System Two to recognize moral injustice. And, to generalize this point, nearly every form of cognitive processing has skill-based components, which contribute alongside – not in the service of and not as realizers of (cf. Carruthers 2009, Frankish 2009, 2016)[19]– states and processes we're inclined to characterize in intellectual terms, as states of System Two. The exercise of skills guides and contributes to all

[19] It's worth commenting on the contrast between the view that the mind is flattened from above, on the one hand, and realization-based views, on the other. Realization is a relation between levels, and thus I reject the proposals that subpersonal reasoning realizes personal-level reasoning (Frankish 2009, 91; *cf.* Frankish 2016) and that System One processes realize System Two processes (Carruthers 2009, 109). Proper qualification is in order, however. Many of the processes normally recognized as subpersonal are, plausibly, realized by other subpersonal processes. Therefore, although the states normally thought to be personal are, on my view, on the same level as many states typically thought to be subpersonal, all of those states (the states erroneously supposed to be at a personal level as well as certain coarse-grained states widely recognized to be subpersonal) are realized by further (more fine-grained) subpersonal states. Another way to approach these issues, perhaps a better way in the end, is to take all of the states in question to be at the same level and to set aside talk of the realization-relation, thinking instead in terms of abstraction or generalities pertaining to a single level.

human cognition in a manner for which we should be most grateful. If it were not for the successful social sensitivity of System One – the automatic, intuitive recognition of duties to our families and communities – societies would not exist today. Were it not for the incredible success of System One at processing language, we would be in no position to talk explicitly about the best way to represent syntactic patterns in a language-manipulating bit of artificial intelligence.

I do not mean to argue that System One is always right with regard to moral or any other matters. Rather, it's to argue, however briefly, such pathologies as System One contributes to should not deflate entirely the regard in which I have recommended that we hold System One. It is largely responsible for very many praiseworthy and impressive products and activities in the domains of art, music, philosophy, sport, reading comprehension, mathematics (for instance, by its contribution to creativity in proofs), conversation, counseling, scientific experimentation, pattern extraction, and moral wisdom. To the extent that System Two can extract and encode helpful messages from those successful activities, various important possibilities are realized. Firstly, System One acts as trainer to System Two, the converse of what was laid out in the first part of this section. Secondly, information encoded in System Two can be fed back into System One, by the training process articulated earlier, which allows for a virtuous feedback loop, as process by which the two systems train up each other over time (despite the fact that this process can go wrong and that, under certain circumstances, it might be best for one system to take primary responsibility for righting the ship). Third, System One can play its role as tutor by tracking differences in System Two's states (perhaps even explicitly representing the content of System Two's states), changing its performance in response – so as to produce even more impressive results – and thereby teaching System Two new things as an indirect effect of its knowledge of System Two.

I conclude that the model of self-improvement through self-knowledge that I've suggested – in which System Two and System One train each other up, partly by tracking the states of the other – is likely to have broad application. Bringing System One and System Two processes into coordination with each other represents a kind of self-improvement that results from each system's knowledge of the other. These systems – or more properly, the variety of cognitive processes that talk of System One and System Two serves as proxy for – are central components of the self-as-cognitive-system; moreover, the coordination of System One and System Two is to be desired and seems properly a feature of the self-as-cognitive-system as a whole. Thus, I have sketched a model of self-improvement based on self-knowledge.

V. Conclusions

What is the entity about which one might have self-knowledge? It would not seem to reside at a distinctive personal level; regardless of whether the idea of such a level is interpreted metaphysically or merely epistemically, a very strong case can be made against its application to the human case. The messiness of the resulting picture – the picture offered by a cognitive science stripped of a personal level – strongly suggests that the object of self-knowledge is the cognitive system in its entirety (or something near enough), including states and activities of its proper parts. The view of the mind as flattened from above and of the self as identical to the entire cognitive system encourages a shift in our conception of self-knowledge. We might take it to be a matter, as much as anything, of the cognitive system's instantiating state-types that systematically track properties of that very cognitive system as a whole or of its proper parts, and that play one or more of the roles we associate of self-knowledge. I have focused on one role of self-knowledge, in particular, that of facilitating self-improvement. I proposed that we conceive of this kind of self-improvement as the result of edifying interaction, and as a thereby effected coordination, between states and processes often thought of as parts of System One and those often thoughts of as parts of System Two.

The primary shortcoming of this picture is, on my view, its excessive liberalism. It treats too much as self-knowledge. There are principled reasons to approach self-knowledge in the way developed, though. Remaining discomfort over excessive liberalism might provide motivation to articulate exactly the objection of excessive liberalism – to see whether it is in fact a high cost (beyond not jibing with the ways we normally talk about self-knowledge or understand it *a priori*). Or, it might give us reason to carve out more precisely the causal-functional roles that the self and self-knowledge should play and to work more diligently to identify, accordingly, attendant restrictions on the states that count as self-knowledge, even on a view of the mind as flattened from above.

Work cited

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

Bermudez, J. L. 2000. "Personal and Sub-Personal: A Difference without a Distinction." *Philosophical Explorations* 1: 63–82.

Bickle, J. 2003. "Empirical Evidence for a Narrative Concept of Self." In G. D. Fireman, T. E. McVay, Jr., and O. J. Flanagan (eds.), *Narrative and Consciousness: Literature, Psychology and the Brain* (Oxford: Oxford University Press), pp. 195–208.

Block, N. 2017. "Unconscious Perception within Conscious Perception." Block's contribution to M. A. K. Peters, R. W. Kentridge, I. Phillips, and N. Block, "Does Unconscious Perception Really Exist? Continuing the ASSC20 Debate," *Neuroscience of Consciousness* 3, 1: 1–11.

Botvinick, M. M., and J. D. Cohen. 2014. “The Computational and Neural Basis of Cognitive Control: Charted Territory and New Frontiers.” *Cognitive Science* 38: 1249–1285.

Buckner, C. 2018. “Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks.” *Synthese*. Online first, 24 September 2018

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

Carruthers, P. 2009. “An Architecture for Dual Reasoning.” In Evans and Frankish 2009, pp. 109–127.

Carruthers, P., and D. M. Williams. ms. “Model-free Metacognition.” Under review.

Chalmers, D. 2004. “Epistemic Two-Dimensional Semantics.” *Philosophical Studies* 118: 153–226.

Chalmers, D. 2012. *Constructing the World*. Oxford: Oxford University Press.

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

Clark, A. 2015. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Clark, A., and A. Karmiloff-Smith. 1993. “The Cognizer’s Innards: A Psychological and Philosophical Perspective on the Development of Thought.” *Mind & Language* 8, 4: 487–519.

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

Cushman, F. 2020. “Rationalization Is Rational.” *Behavioral and Brain Sciences* 43, e28: 1–59. doi:10.1017/S0140525X19001730.

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

Davies, M. 2000b. “Persons and Their Underpinnings.” Philosophical Explorations 3: 43–62.

De Houwer, J., B. Gawronski, & D. Barnes-Holmes. 2013. “A Functional-Cognitive Framework for Attitude Research,” *European Review of Social Psychology* 24, 1: 252–287.

De Vignemont, F. 2014. “A Multimodal Conception of Bodily Awareness.” *Mind* 123: 989–1020.

Dennett, D. C. 1969. *Content and Consciousness*. Abingdon, UK: Routledge.

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

Dewhurst, S. A., Conway, M. A., & Brandt, K. R. (2009). “Tracking the R- to-K Shift: Changes in Memory Awareness across Repeated Tests.” *Applied Cognitive Psychology* 23: 849–858.

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

Drayson, Z. 2014. “The Personal/Subpersonal Distinction.” *Philosophy Compass* 9 (5): 338–346.

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

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

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

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

Flanagan, O. 1994. “Multiple Identity, Character Transformation, and Self-Reclamation.” In G. Graham and G. L. Stephens (eds.) *Philosophical Psychology* (Cambridge, MA: MIT Press), pp. 135–162.

Fodor, J. A. 1974. “Special Sciences (Or: The Disunity of Science as a Working Hypothesis)” *Synthese* 28: 97–115.

Fodor, J. A. 1975. *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: MIT Press.

Fodor, J. A. 1998. *Concepts: Where cognitive science went wrong*. Oxford: Oxford University Press.

Frankish, K. 2009. “Systems and Levels: Dual-System Theories and the Personal-Subpersonal Distinction.” In Evans and Frankish 2009, pp. 89–107.

Frankish, K. 2016. “Playing Double: Implicit Bias, Dual Levels, and Self-control.” In M. Brownstein and J. Saul (eds.), *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology* (Oxford: Oxford University Press), pp. 23–46.

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

Gathercole, S. E., and A. Baddeley. 1993. *Working Memory and Language* (Hillsdale, NJ: Lawrence Erlbaum).

Gawronski, B., and G. 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, T. 2008*a*. “Alief and Belief.” *Journal of Philosophy* 105: 634–663.

Gendler, T. 2008*b*. “Alief in Action (and Reaction).” *Mind & Language* 23, 5: 552–585.

Gertler, B. 2011. *Self-Knowledge*. London: Routledge.

Giere, R. 2006. “The Role of Agency in Distributed Cognitive Systems.” *Philosophy of Science* 73: 710–719.

Gigerenzer, G. 2000. *Adaptive Thinking*. Oxford: Oxford University Press.

Goldman, A. 1979. “What Is Justified Belief?” In G. Pappas (Ed.), *Justification and Knowledge* (Boston: D. Reidel), pp. 1–25.

Holton, R. 2016. “Review of Andy Clark, *Surfing Uncertainty*.” *Times Literary Supplement* 7 October 2016.

Hornsby, J. 2000. “Personal and Sub-Personal: A Defence of Dennett’s Early Distinction.” *Philosophical Explorations* 3 (1): 6–24.

Hurley, S. L. 1998. “Vehicles, Contents, Conceptual Structure, and Externalism.” Analysis 58 (1): 1–6.

Ismael, J. T. 2014. “On Being Someone.” In A. R. Mele (Ed.), *Surrounding Free Will: Philosophy, Psychology, Neuroscience* (Oxford: Oxford University Press), pp. 274–297.

Karmiloff-Smith, A. 1992. *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Kim A., and L. Sikos. 2011. “Conflict and Surrender during Sentence Processing: An ERP Study of Syntax-Semantics Interaction.” *Brain and Language* 118: 15–22.

Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.
Levy, N. 2016. “‘My Name is Joe and I’m an Alcoholic’: Addiction, Self-knowledge and the Dangers of Rationalism.” *Mind & Language* 31, 3: 265–276.

Lyons, J. 2016. “Unconscious Evidence.” *Philosophical Issues* 26 (*Knowledge and Mind*): 243–261.

Machery, E. 2009. *Doing without Concepts*. New York: Oxford University Press.

Mandelbaum, E. 2015. "Attitude, Inference, Association: On the Propositional Structure of Implicit Bias." *Noûs* 50, 3: 629–658.

Marr, D. 1982. *Vision*. New York. W H. Freeman.

McDowell, J. 1994. The content of perceptual experience. *Philosophical Quarterly* 44, 175: 190-205.

Miracchi, L. 2017. "Perception First." *Journal of Philosophy* 114, 12: 629–677.

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

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

Papineau, D. 2001. "The Rise of Physicalism." In C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents* (Cambridge: Cambridge University Press), pp. 3–36.

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

Pylyshyn, Z. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Quilty-Dunn, J. 2020. "Rationalization Is Irrational and Self-serving, but Useful." *Behavioral and Brain Sciences* 43, e28: 30–31 doi:10.1017/S0140525X19001730.

Quine, W. V. 1969. "Epistemology Naturalized." In Quine, *Ontological Relativity and Other Essays* (New York: Columbia University Press), pp. 69–90.

Rey, G. 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.

Robins, S. K. 2017. "Memory Traces." In S. Bernecker and K. Michaelian (Eds.) *Routledge Handbook of the Philosophy of Memory* (Abingdon, UK: Routledge), pp. 76–87.

Rowlands, M. J. 2009. "Extended Cognition and the Mark of the Cognitive." *Philosophical Psychology* 22, 1, 1–19.

Rowlands, M. J. 2010. *The New Science of the Mind: From Extended Mind to Embodied Phenomenology* (Cambridge: MIT Press).

Rupert, R. D. 2009. *Cognitive systems and the extended mind*. Oxford: Oxford University Press.

Rupert, R. D. 2010. Extended cognition and the priority of cognitive systems. *Cognitive Systems Research* 11: 343–56.

Rupert, R. D. 2011*a*. "Review of Mark Rowlands, *The New Science of the Mind: From Extended Mind to Embodied Phenomenology*." *Notre Dame Philosophical Reviews* 2011.03.35 (March 31, 2011)

Rupert, R. D. 2011*b*. "Embodiment, Consciousness, and the Massively Representational Mind," Philosophical Topics 39, 1: 99–120.

Rupert, R. D. 2013. "On the Sufficiency of Objective Representation." In U. Kriegel (ed.), *Current Controversies in Philosophy of Mind* (New York: Routledge), pp. 180–196.

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

Rupert, R. D. 2016. "Embodied Concepts, Conceptual Change, and *A Priori* Knowledge; or, Justification and the Ways Life Can Go," *American Philosophical Quarterly* 53, 2 (April, 2016): 169–192.

Rupert, R. D. 2018. "The Self in the Age of Cognitive Science: Decoupling the Self from the Personal Level." *Philosophic Exchange* 47: 1–36.

Rupert, R. D. 2019. "What Is a Cognitive System? In Defense of the Conditional Probability of Co-contribution Account." *Cognitive Semantics* 5: 175–200.

Schacter, D. L. 2012. "Constructive Memory: Past and Future." *Dialogues in Clinical Neuroscience* 14, 1: 7–18.

Schechtman, M. 2007. "Stories, Lives, and Basic Survival: A Refinement and Defense of the Narrative View." *Royal Institute of Philosophy Supplement* 60: 155–178.

Schechtman, M. 2011. "The Narrative Self." In S. Gallagher (Ed.), *The Oxford Handbook of the Self* (Oxford: Oxford University Press), pp. 394–416.

Schroeter, L. 2004a. "The Limits of Conceptual Analysis," *Pacific Philosophical Quarterly* 85, 4: 425–453.

Schroeter L. 2004b. "The Rationalist Foundations of Chalmers's 2-D Semantics." *Philosophical Studies* 118: 227–255.

Schroeter, L. 2006. “Against *A Priori* Reductions,” *Philosophical Quarterly* 56, 225: 562–586.

Shea, N. 2013. “Neural Mechanisms of Decision-Making and the Personal Level.” In K.W.M. Fulford, M. Davies, G. Graham, J. Sadler, G. Stanghellini, and T. Thornton (eds.) *Oxford Handbook of Philosophy and Psychiatry*. Oxford: Oxford University Press, pp. 1063–1082.

Shea, N. 2018. *Representation in Cognitive Science*. Oxford: Oxford University Press.

Stich, S. P. 1996. *Deconstructing the Mind.* New York: Oxford University Press.

Velleman, D. 2005. “The Self as Narrator.” In J. Christman and J. Anderson (eds.), *Autonomy and the Challenges to Liberalism: New Essays* (Cambridge: Cambridge University Press), pp. 56–76.

Wilson, M. 2002. “Six Views of Embodied Cognition.” *Psychonomic Bulletin and Review* 9: 625–636.

Wilson, T. D., 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.