Synthese (2021) 198:9243–9272
https://doi.org/10.1007/s11229-020-02631-1

Epistemic value in the subpersonal vale

J. Adam Carter[1] · Robert D. Rupert[2]

Received: 31 May 2019 / Accepted: 14 March 2020 / Published online: 2 April 2020

Abstract
A vexing problem in contemporary epistemology—one with origins in Plato's *Meno*—concerns the value of knowledge, and in particular, whether and how the value of knowledge exceeds the value of mere (unknown) true opinion. The recent literature is deeply divided on the matter of how best to address the problem. One point, however, remains unquestioned: that if a solution is to be found, it will be at the personal level, the level at which states of subjects or agents, as such, appear. We take exception to this orthodoxy, or at least to its unquestioned status. We argue that subpersonal states play a significant—arguably, primary—role in much epistemically relevant cognition and thus constitute a domain in which we might reasonably expect to locate the "missing source" of epistemic value, beyond the value attached to mere true belief.

Keywords Subpersonal cognition · Epistemic value · Virtue epistemology · Philosophy of cognitive science

1 Epistemic value and the swamping problem

The past decade or so has witnessed a 'value turn' in mainstream epistemology (Riggs 2008). Of particular interest have been matters connected to epistemic value, value

An earlier version of this paper was presented by R.D.R. to Albert Newen's research group at Ruhr University, Bochum, at the University of Edinburgh, and at a Morris Colloquium at the University of Colorado, Boulder. R.D.R. thanks members of all three audiences for useful and simulating feedback, as well as Carrie Figdor, Bob Pasnau, Heather Demarest, and Raul Saucedo, who commented on a draft of the manuscript.

✉ J. Adam Carter
adam.carter@glasgow.ac.uk

Robert D. Rupert
robert.rupert@colorado.edu

[1] University of Glasgow, Glasgow, UK

[2] University of Colorado, Boulder, Boulder, USA

arising from the existence of distinctively intellectual goods, in contrast to, for example, moral or aesthetic goods (Kvanvig 2003; Pritchard 2007, 2011; Haddock et al. 2009; Baehr 2009). The investigation of epistemic value has generated various puzzles, one of which owes its existence partly to a compelling piece of received wisdom: that the epistemic value of knowledge outstrips that of mere true belief[1]; it may, for instance, be epistemically valuable to believe some true proposition <*p*> via a fortunate guess, but it is even more epistemically valuable to *know* that <*p*>. A subject who, for example, works stepwise through a proof of logical theorem T_{1L} and, as a result, correctly affirms T_{1L}'s status is in a better epistemic state than the person who unthinkingly spits out 'yes, it's a theorem' as a mere guess or only because, say, that person likes the font in which T_{1L} is set. And this is the case, to reiterate, even though *both* subjects end up with a true belief about T_{1L}'s status as a theorem.

This apparent truism makes mischief in the following way. Consider what would seem to be an equally plausible principle: whatever we say about the value of knowledge must be consistent with what we say about the *nature* of knowledge (Kvanvig 2003). If we accept that the epistemic value of knowledge exceeds that of mere true belief, we should not endorse an analysis of knowledge foreclosing that very possibility; rather, our analysis of knowledge must comport with an account of its constituents (e.g., justification or warrant) such that, plausibly, being justified (or warranted) adds value to what would otherwise be a merely true belief. But, some influential accounts of knowledge have struggled to satisfy this conditional constraint.

Take, for example, a straightforward process-reliability account of knowledge, according to which knowledge is type-identical to reliably produced true belief (Goldman 1979; Olsson 2007). If the process reliabilist maintains that the epistemic value of knowledge exceeds that of mere true belief, it is incumbent upon her to demonstrate just *how* the epistemic value of reliably produced true belief could exceed the value of mere (that is, not reliably formed) true belief. Here the reliabilist runs into trouble. Consider Zagzebski's telling comparison (Zagzebski 2003): an already good-tasting cup of coffee takes on no additional gustatory value simply in virtue of its being the product of a machine that reliably produces good-tasting cups of coffee; likewise, it is unclear how a true belief would become *additionally* epistemically valuable if turns out that the belief is not only true, but also the product of a reliable belief-forming process.[2] It seems that the epistemic value of truth "swamps" the value of reliability (Kvanvig 2003, pp. 47–48). But—and here's the bad news for reliabilism—if the epistemic value of truth swamps the epistemic value of reliability, then it is false that the epistemic value of reliably formed true belief exceeds the epistemic value of mere (unreliably formed) true belief. Thus, the reliabilist's account of the nature of knowledge appears not to be consistent with the assumption that the epistemic value of knowledge exceeds that of mere true opinion.

And though this swamping problem (Kvanvig 2003, 2010; Zagzebski 2003; Jones 1997; Pritchard 2009b; Swinburne 1999; Sylvan 2017) is often treated as an objection specifically to (at least standard forms of) reliabilism, it threatens other analyses of

[1] In Plato's *Meno*, Socrates asserts (uncharacteristically) that he knows this, if he knows anything at all. See Kvanvig (2003, Ch. 1) for discussion.

[2] Kvanvig (2003) makes a parallel point in terms of the value of chocolate.

knowledge in like fashion (Kvanvig 2003, 2010; Pritchard 2009a, b): in the case of *any* account of knowledge which aims to uphold the plausible assumption that the epistemic value of knowledge exceeds that of mere true belief, whatever conditions must, according to that account, be satisfied by knowledge must themselves be such that their satisfaction adds epistemic value to a true belief. And, virtually all otherwise promising theories of knowledge (or warrant, or justification) seem to flounder in the face of this challenge.[3]

Responses to the swamping problem typically take one of three forms: *validationist, fatalist*, or *revisionist* (Pritchard 2009a, pp. 19–20). Validationists (e.g., Greco 2010; Goldman and Olsson 2009; Olsson 2007, 2011) argue that, when knowledge is understood properly, the epistemic value of knowledge is not swamped by the value of true belief. By contrast, *fatalists* (e.g., Baehr 2009; Ridge 2013) argue that we circumvent the swamping problem by rejecting the underlying intuition about the value of knowledge. Maybe, as this line of thought goes, knowledge is not as valuable as we initially thought it was. Finally, *revisionists* (e.g. Kvanvig 2003, 2010; Pritchard 2009a, b, 2010; Riggs 2009) agree with fatalists that knowledge lacks a distinctive epistemic value, one not shared by mere true belief, while claiming that it is on closer inspection something *else*—typically understanding—which possesses the distinctive sort of epistemic value mistakenly attributed to knowledge.

2 The swamping problem and the personal/subpersonal distinction

2.1 The orthodox view

Despite deep disagreement about how best to solve the swamping problem, all parties seem to agree where to look in conceptual space: if there is a solution, it is to be found at the *personal level*. What is the personal level? In broad strokes, it is the realm of states of persons "as such, as experiencing, thinking subjects and agents" (Davies 2000a, p. 88, summarizing Dennett's original thought when introducing the idea of a distinctively personal level of explanation or description).[4] Although the distinction between the personal and subpersonal levels is frequently taken for granted, the details could use some sorting out. Such sorting out is not our project here. Thus, in the remainder, we will simply follow precedent and take accessibility to consciousness—typically operationalized as reportability—and suitability for appearance in folk psychological and rationalizing explanations as the primary marks of personal-level status.[5]

[3] Pritchard (2010) has suggested that, among existing contenders, robust virtue epistemology (e.g. Sosa 2007; Greco 2010) boasts resources most likely to vindicate the value of knowledge in the face of the swamping problem. However, at least as Pritchard sees it, robust forms of virtue epistemology are materially inadequate. And, perhaps more germane to present purposes, virtue-theoretic proposals are not sufficiently informative when pitched at the personal level, a concern we develop in more detail below.

[4] Dennett introduced the distinction (1969, p. 93) as a difference in styles of explanation (or between sets of explanatory vocabularies). In our discussion, we focus on the corresponding ontological questions—of personal-level states, properties, or processes – as is frequently done in the contemporary literature.

[5] Davies (2000a, pp. 88–90, 2000b, p. 46) focuses on these characteristics of personal-level states and processes, as do Shea (2013, pp. 1064–1065) and Frankish (2009, pp. 90–91); on the messiness of the personal-subpersonal distinction, see Drayson (2012, 2014).

The standard formulation of the swamping problem presupposes a contrast between knowing and merely truly believing, and both belief and knowledge are widely thought to be states of whole persons as such; one's beliefs (neat and as components of knowledge states) are typically available to introspection, and even more straightforwardly, they are clear candidates for inclusion in rationalizing and folk psychological explanations of one's actions. Moreover, these personal-level states are thought by many to have a distinctive sort of content; in fact, it is sometimes claimed that the kind of content had distinctively by personal-level states is the only genuine form of content and that only this sort of content has such genuine epistemic properties as *carrying justificatory force* (McDowell 1994a, b). Generally, the consensus holds that personal-level states—beliefs, memories, perceptions—and justification-relevant relations between them provide the subject matter of epistemology and thus that, if a solution is to be found to the swamping problem, it will be found there—in the nature of personal-level states and relations between them or between their contents.

We take exception to this orthodoxy, or at least to its unquestioned status. After all, the further cognitive science has progressed, the greater the extent to which its results—some of which we will describe in Sects. 3 and 4—have marginalized the personal level in its accounts of human behavior (Nisbett and Wilson 1977; Wegner 2002; T. D. Wilson 2002; Lau et al. 2007; Haybron 2007; Schwitzgebel 2008; Alfano 2013; Harman 2000; Doris 2002; Gendler 2008). Given the state of the empirical evidence, then, it makes good sense for epistemologists to look to the subpersonal level. We contend that such exploration holds promise, that subpersonal states and their content[6] play a significant role in epistemically relevant cognition—*particularly in cases in which there is no substantive model of what one might think should be the corresponding personal-level state or process*—and thus constitute a domain in which we might reasonably expect to locate the (or, at least *a*) missing source of epistemic value, beyond the value attached to mere true belief.

2.2 Methodological remarks

Two preliminary comments about our argumentative strategy are in order. First, we do not press the following, relatively trivial point: that all personal-level states or processes are grounded in, enabled by, or realized by physical states or processes, or supervene on physical states or processes (of the brain, it is typically thought). A broad consensus in philosophy of mind and epistemology accepts some form of physicalism—typically cashed out as a claim about realization or supervenience—and we do not intend merely to endorse that consensus. We have a narrower point in mind. We claim that, in some cases, personal-level states have epistemic value the source

[6] Although we emphasize the role of so-called subpersonal-level states and take the content of such states to be relevant to questions about epistemic value, we do not commit ourselves to a distinctive form of content at the subpersonal level. It might be that the content of the relevant subpersonal states is of the same sort as the content of supposed personal-level states, but that the state-types are different (*cf.* Heck 2000, which makes a parallel point regarding the conceptual-nonconceptual distinction). That is to say, we might solve the swamping problem by moving the discussion to the subpersonal level, though the content found at that level might or might not differ in its nature from the kind of content typically supposed to appear at the personal level.

of which seems mysterious viewed from the standpoint of the personal level. In such cases, there is no straightforwardly identifiable personal-level process or set of relations that might account for the value of such states. Yet, when one takes the subpersonal perspective, the source of epistemic value comes more clearly into view. The matter might be best seen as involving a structural mismatch: in the cases at issue, a minimal amount of structure appears at the personal level, while a more richly structured process appears at the subpersonal level; as a result, the attempt at personal-level analysis falls short for want of explanatory resources, while, in contrast, subpersonal-level structure and processes offer plentiful explanatory resources, many of them relevant to questions of justification, or so we will argue.

Second, although our arguments in Sects. 3 and 4 presuppose a relatively demanding *criterion of adequacy* for a validationist response, it is not as demanding a criterion as some might wish. We tend to agree with Duncan Pritchard (2013, p. 12) that ‘in general and all other things being equal, we desire to be knowers as opposed to being agents who have mostly true beliefs but lack knowledge’. Showing that this desire is not misguided does not require showing that, in every instance (actual or possible), knowledge that *P* is more valuable than mere true belief that *P*. If a successful validationist response were to require meeting the strong (we believe, unduly strong) demand that *all possible* items of knowledge have a value that exceeds the value of their corresponding true belief tokens, then the arguments we develop below would likely come up short. However—as Kvanvig (2003) himself has argued at length—it’s not at all obvious that there is, at the end of the day, *any* way of defending a validationist response if such a strong modal criterion of adequacy is assumed. And, this point gains further traction when (following Sosa 2000, see also David 2001; Lynch 2009) one reflects on the tension between the strong modal criterion of adequacy and the apparent ‘pointlessness’ of attaining knowledge of trivial truths (e.g., truths about the number of grains of sand on an arbitrary section of a beach). The reader should bear in mind, however, that we do not aim to satisfy only the exceptionally weak demand that in *some* actual cases, knowledge that P has more epistemic value than the mere true belief that P. Rather, we aim to satisfy an intermediate and practically relevant demand: to show that in a wide range of actual cases involving human subjects, knowledge that P has more value than the mere true belief that P.

With these points in mind, here is the plan. Section 3 relates our shift in perspective—from the personal level to the subpersonal level—specifically to extant, proposed personal-level solutions to the swamping problem on behalf of the process reliabilist (Olsson 2007) and virtue epistemologist (Greco 2010), arguing that such proposals show more promise when recast at least partly in subpersonal terms and supplemented accordingly. In Sect. 4, we move into entirely uncharted territory, by proposing and defending two arguments directly in support of a subpersonal validationist solution to the swamping problem, a solution we defend as legitimate with reference to the weaker criterion of adequacy that Pritchard seems to have in mind and which isn’t predicated upon what are perhaps overoptimistic assumptions about what it *is* that should be validated.

3 Subpersonal transformations of personal-level proposals

In this section, we argue, in 3.1, that one extant reliabilist attempt to solve the swamping problem—viz. Olsson's (2007) argument from increased practical value—faces what appear, from the empirical standpoint, to be insurmountable hurdles. Upon consideration of further empirical results, we conclude that Olsson's central claim—that justification confers value by conferring stability—has more plausibility when one focuses on subpersonal processes the structural images of which do not appear at the personal level. In Sect. 3.2, we consider what is taken to be the most promising personal-level response to the swamping problem currently on offer—viz. the solution offered by (robust forms of) virtue epistemology (e.g., Greco 2010)—and argue that a subpersonal variation on this approach has more to recommend it.

3.1 Process reliabilism

Erik J. Olsson (2007) argues that "reliabilist knowledge promotes successful action over time…[because]…reliabilist knowledge promotes stability and…stability is conducive to successful action over time" *(ibid*., 349). According to Olsson, when an unreliable process produces a (mere) true belief that *P*, the very unreliability of that process will likely undercut or neutralize, eventually, that belief's potential to contribute to successful action. A given subject deploys a given mechanism or runs a given process-type (we treat these as equivalent for present purposes) repeatedly over the course of her life. In the cases in question, the process-type is, by hypothesis, unreliable; thus, the preponderance of later applications of it—that is, those that occur after the time at which the application of the process led to the fixation of the subject's true belief that *P*—will yield false beliefs, which will likely lead to unsuccessful action. As Olsson sees things, subjects track the sources of their belief, recording which processes produce which beliefs as well as the rate of past success and failure of various processes to produce beliefs that lead to effective action. Thus, a subject who continues to use the process in question will subsequently doubt or reject *P*; given feedback from the world, the subject will detect the falsity of the outputs of the majority of later applications of the process that produced *P*, which results will call into question *P* itself (even though *P* is, in fact, true), thereby robbing the true belief that *P* of what would have been its contributions to successful behavior—presumably because the subject abandons, or at least take a highly qualified attitude toward, the belief that *P*, and thus does not act on it.[7]

[7] This work builds on the simpler idea (Goldman and Olsson 2009) that having a reliably produced true belief is better than having a true belief produced by an unreliable process, because one's having a reliably produced true belief probabilifies one's having true beliefs in similar circumstances in the future. While this may be correct, it does not seem to increase the value of any individual belief. A true belief's having been produced by a reliable process entails the presence of a valuable tool in the subject, a tool such that, if the subject continues to possess and deploy it, it will produce a preponderance of true beliefs in the future. It is not clear, however, why the value of the possession of that tool would increase the value of a given belief produced by it, beyond the value the belief has in virtue of its truth. Instead, a belief's having been produced by a reliable process seems to be merely an indicator that the subject possesses a tool to produce true beliefs reliably (in certain kinds of circumstances); the relational fact of a belief's having been produced by a reliable process would seem to confer only a diagnostic role on that belief.

To be clear, to the extent that Olsson's tack succeeds, it does not provide what many would want from a response to the swamping problem: an account of why a justified true belief that qualifies as knowledge has more *final* (i.e., noninstrumental) value than a mere true belief. Rather, it focuses on a particular kind of instrumental value, viz. the value that a justified true belief has in virtue of its being likely, itself, to continue to contribute to successful action. Nevertheless, the account does generally make sense of the intuition that justified true belief is more valuable than (mere) true belief, given the incredible importance of successful action in human lives.

Let's assume, for the sake of argument, that a proposal along the lines of Olsson's holds promise. Whence does it draw its explanatory power, the personal level or the subpersonal level? At what level does the connection between justification and value-qua-stability appear? As noted above, Olsson supposes that agents generally record which beliefs were produced by which mechanisms and keep a running success rate of each mechanism. If the agent were not successfully deploying such record-keeping abilities, she would not infer the likely falsity of *P* from later failures of the *P*-producing process to yield beliefs that support successful action. So far as we can tell, then, this is meant to be a proposal about personal-level states. In Olsson's words, his proposal "requires that the agent maintain a mental record, a record in her mind, of how beliefs were acquired" (Olsson 2007, p. 352).

A fleshed-out version of Olsson's story would seem to require that, relative to each use of a given belief-forming mechanism, the subject accurately encode, not just *that* the mechanism produced the belief in question, but also the context of such production, represented at correct level of specificity. Too often, a mechanism that has a weak track record in a context described at one level of specificity (use of vision while beneath the surface of the water in a naturally formed lake) has a stronger track record relative to a more inclusive set (use of vision *simpliciter*), or vice versa—points familiar from discussions of the Generality Problem for process reliabilism.[8]

Assume a given true belief *P* was produced by a mechanism in a context that would be appropriately individuated, for the purpose of determining *P*'s level of justification, at a fine grain, and that relative to such a context, the mechanism is in fact unreliable. Let us say, too, that the subject mistakenly individuates, in her record-keeping process, the context in question in a coarser-grained way relative to which the mechanism in question is, in fact, reliable. In this case, the subject will *not* weed out the unjustified belief that *P*, because she will treat the contexts of the application of the mechanism in question in a coarser-grained fashion and will not come to see—on the basis of negative feedback from the world in the relevant finely individuated contexts—the mechanism that originally produced the belief that *P* as unreliable. If this is a relatively common phenomenon—if the subject doesn't identify and record contexts at the correct grain—the subject's various unjustified true beliefs may well persist and continue to contribute to successful behavior, *contra* Olsson's prediction; the subject will think of *P* as the result of the operation of a reliable mechanism (the visual system applied to the general spatial layout, for example), which generally gets things right, instead of seeing the mechanism as being applied in more finely individuated context (the visual system as applied in poor light to objects at a distance moving quickly), and

[8] For a recent influential discussion of this problem, see Lyons (2019).

thus won't abandon her unjustified belief that P.[9] Similarly, without accurate records of the sort in question, the subject may well treat a justified true belief as unjustified and abandon it in accordance with Olsson's schema. Thus, absent a commitment to the reasonably accurate tracking and recording of justification-relevant contexts of belief-formation, one should doubt that Olsson's schema identifies the distinctive source of value attached to justification.

How plausible is it, then, that human subjects track, at the personal level, the output of belief-forming mechanisms in contexts? Below we survey the relevant empirical literature, but let us be clear, from the outset, about the sort of evidence we should look for on Olsson's behalf. It should show that, at the personal level, (a) subjects track the sources of their beliefs (and the contexts of their formation—take this as read in what follows), (b) subjects do so reasonably accurately, (c) subjects use that information to calculate reasonably accurate track records of their various belief-producing mechanisms, and (d) subjects bring the results of those calculations to bear on commitments to past products of their belief-producing mechanisms (for example, to bear on the judgment that a past doxastic output of a given mechanism is likely to be true) and adjust accordingly. Moreover, given that Olsson's account of how justification adds value (by creating stability) is pitched at the personal level, we should expect subjects to be able to report accurately on the states and processes in question in connection with (a–d). This sets a very high bar, to be sure, and, perhaps unsurprisingly, the empirical literature runs in the opposite direction, supporting at least a moderately pessimistic view about every one of these desiderata and, more importantly, a firm skepticism concerning their joint satisfaction. The literature in question is enormous, but we shall, in what follows, convey a sense of the obstacles faced by Olsson's personal-level account.

Consider first the so-called illusory-truth effect (Hasher et al. 1977; Dechêne et al. 2010). Mere exposure to sentences increases the likelihood that subjects will judge them to be true when asked later about them. This occurs even in cases in which subjects know better, that is, even when they have stable beliefs that contradict the information to which they're being exposed experimentally (Fazio et al. 2015). The leading explanation of this phenomenon appeals to a more general and well-established subpersonal construct, ease of processing, in this case created by previous exposure (Begg et al. 1992).[10] Further results reinforce this hypothesis: merely setting (unfa-

[9] This kind of concern is not merely theoretical. For instance, the sort of source information that some languages encode syntactically marks sources only at a very coarse grain—distinguishing between such categories as having been acquired by testimony or having been observed first-hand (Tosun et al. 2013).

[10] Some such patterns of judgment are accompanied by a reported sense of confidence (a "sense of knowing") the strength of which reliably increases with, e.g., the number of past exposures to the sentence in question. We should not, however, take such a reportable sense of knowing to indicate that the processes producing the judgments concerning truth or validity—processes exhibiting ease-of-processing effects, for instance—operate at the personal level. Such a criterion would be too weak, for it would miscategorize paradigmatically subpersonal-level processes as personal. Consider textbook approaches to speech processing, the best known of which is Chomskyan generative grammar. It's widely accepted that subjects' conscious sense of what's grammatical and what's not—their "linguistic intuitions"—provides important data in theory construction in this domain. But, the theories so constructed involve processes (computational sensitivity to the presence of 'wh'-traces, that no subject can report, for instance) that are subpersonal if any cognitive processes are. Similarly, consider processes operative in face recognition. A reportable sense of familiarity with a given face may correspond with the face's having been categorized correctly.

miliar) sentences in an easier-to-process font increases the probability (over sentences set in a more difficult to read font) that subjects will judge them to be true after a single previous exposure (Reber and Schwarz 1999). In such cases, subjects simply are not attending accurately to the sources of their beliefs or to the mechanisms producing them. If they were aware, at the personal level, that a mere ease-of-processing mechanism were producing the beliefs in question—responding only to, for instance, previous exposure to written text with no attached credibility—subjects would, presumably, not make the judgments they do.

The illusory truth results are by no means outliers. Empirical work documents various kinds of cases in which subjects fail to track the sources of their beliefs and mechanisms that produced them (see Marsh et al. 2008, for a review). In a well-known list-learning paradigm, for example, subjects are exposed to a list of semantically interrelated words and asked later about words that would have "fit" onto the list (but weren't on it); subjects frequently judge that these words were listed. Subjects seem to have both a false belief and to make a false judgment about the source of that belief, thinking they heard or read the word when it was instead self-generated (by a subpersonal process of semantic association). And, in some versions of such experiments, the personal-level record is strikingly corrupted; subjects report phenomenological experience—a rich episodic memory—of, for instance, the experimenter having read the nonlist word aloud, even though it is only a lure and was not in fact read aloud (Roediger and Gallo 2005; see Geraci and Franklin 2004 for cases in which subjects are misled by nonsemantic linguistic relations). In eyewitness suggestibility experiments, subjects report having witnessed what are actually false details that experimenters have in one way or another exposed the subjects to after the witnessing of the actual event in question; subjects confuse testimony-based belief-formation (verbal or in print) for first-person observations made at, for example, the scene of a car accident (Loftus 1979). A different line of research shows that déjà vu can be induced experimentally, by means of mere exposure, which again seems to be a mistake about sources (Brown and Marsh 2010). Subjects are susceptible to the false fame effect, incorrectly categorizing faces as being those of famous people, as result of mere past exposure to said faces in experimental settings (Jacoby et al. 1989a, b). And, subjects are more likely to choose the wrong subject from a line-up when they've seen that person's face in a book of "mug shots" prior to viewing the line-up (Brown et al. 1977). Subjects also make source errors when recalling the factors that influenced their decisions, in a way that systematically supports decisions made (Mather et al. 2000), attributing, for example, positive features to the option they chose, even when those positive features were actually attached to the option not taken. In addition, the effects of social contagion powerfully distort memory of sources (Meade and Roediger 2002; Barnier et al. 2008). And bear in mind that in every one of these cases—as well as in the cases of memory-related results discussed in the remainder of this subsection—the measures used involve at least some (and often exclusively) personal-level judgments.

Footnote 10 continued
But, this personal-level sense of familiarity puts subjects in no position to report on the structure of the process that extracts the fine-grained geometrical features of faces and that produces successful acts of face-recognition—by, for instance, calculating the distance, in a multi-dimensional feature space, of the face currently represented from various exemplar faces (that is, standing representations of known individuals' faces). That process is clearly subpersonal.

Results on source credibility should be especially troubling to Olsson. In some cases, subjects' beliefs about the credibility of the source of an individual piece of information significantly affect their judgments: when they believe that a piece of information is from a credible source, they're significantly more likely to judge it true. But, even then, they have such poor memory for the actual sources of individual memories (Begg et al. 1992, p. 452) that, statistically speaking, the actual credibility of the source is not correlated with the subjects' pattern of endorsements (Henkel and Mattson 2011, p. 1708). Moreover, subjects continue to categorize as true approximately 50% of statements they believe (rightly or wrongly) to be from a noncredible source (Begg et al. 1992, pp. 451–453). The picture that emerges is of subjects who are not completely insensitive to considerations of credibility, but who have (1) mediocre source memory, (2) often make poor use of credibility information they have, and (3) who continue to make widespread errors concerning other matters to do with sources. Considering the multi-step, statistical nature of the procedure Olsson demands, such shortcomings compound. As a result, there's no reason to believe of any particular mechanism in a given subject, that the subject will have a sufficiently accurate record of the performance of that mechanism and will take it into consideration in the formation of new beliefs or the continued endorsement of belief previously produced by that mechanism.

To be fair, some experimenters have gone to significant lengths to try to warn subjects about credibility or to get them to use information about credibility to mitigate the formation of or reliance on false memories (Echterhoff et al. 2005; Henkel and Mattson 2011; Chambers and Zaragoza 2001; Meade and Roediger 2002; Begg et al. 1992). These efforts are not wholly without results, but neither do they instill much confidence in human abilities or tendencies. To the extent that such warnings have salutary effects (reducing eyewitness suggestibility effects, for example), the effects are weak, depend on the specific choice of wording (the warning should be given in the indicative, not the subjunctive), and on the timing of the warning (warnings are more effective if they're provided before the "false testimony" not after). This last point, in particular, stands in direct tension with Olsson's framework. On Olsson's account, warnings that a belief-producing mechanism "lacks credibility" come only *after* the mechanism's production of the original belief in question; in the case of a merely true belief that *P*, the belief lacks stability because *later on* it becomes apparent to the subject that the mechanism that produced it is producing further beliefs that fail to guide successful action. And, even in the specific cases in which the warnings work well, for instance, eliminating the effects of misinformation, this amounts only to the subjects' treatment of planted information in the same way as subjects treat new information (that is, stimulus items first introduced at the time of later memory tests or retests). But, given how badly subjects perform on newly introduced material—for example, given the high rate at which subjects falsely claim that newly introduced items were present in the earlier study material—Olsson will find little consolation here. It would be one thing if prewarnings could somehow get subjects to track sources reliably enough to engage in the personal-level feats of memory required by Olsson's framework. But it's another to be told merely that prewarnings can prevent subjects from being misled by previous-exposure effects, but that subjects in these cases nevertheless do not very reliably judge whether they've *ever* been exposed to a given stimulus item (that is, whether the current experience has a source in memory at all). Notice, too, that the beneficial effects are relatively short-lived (Chambers and Zaragoza 2001, p. 1122), which further undermines Olsson's picture. For on his framework, the "warning of unreliability" is likely to come *well* after the fact—when, significantly later, the mechanism in question produces further beliefs that do not support successful action. Even if the subject is given a prewarning that alerts her to the potentially misleading ways of a given source, if, before long, she stops paying attention to said warning, the benefits of circumspection will have been lost by the time she must, on Olsson's account, notice that a mechanism that produced the merely true belief that P is now producing beliefs that fail to support successful action.

Perhaps it is unsurprising, on general theoretical grounds, that subjects lack the capacity for reportable, detailed memory that Olsson's framework requires. Marcia K. Johnson and collaborators (Johnson et al. 1993) developed the leading theory of the monitoring of sources of information and memories. Put simply, Johnson's source-monitoring framework grounds subjects' ability to identify the source of a given memory in the ability to associate the content of the memory itself with various cues and features of the context in which the memory was formed. Viewed in that light, the various results reviewed above may seem unsurprising, given the general remember-to-know (or, R-to-K) shift (Barbar et al. 2008; Dewhurst et al. 2009) exhibited by human memory: the general tendency for information about the specific circumstances to be lost over time and for the supposed knowledge to be represented context-free. This distinction parallels (and may be largely coextensive with) the widely made distinction between episodic and semantic memory (Tulving 1972) and the tendency of memories for general information about the world to shift from episodic form—replete with details concerning the context in which the information was acquired—to the semantic form, which encodes the information itself stripped of such details. To the extent that recall includes detailed information, it is typically reconstructed by a process that allows, relatively easily, for error to creep into the representation of those details.

In sum, the breadth and the depth of subjects' personal-level mistakes—their failure to mark sources at all, their failure to mark sources accurately, their failure to use accurate information that they have—puts paid to Olsson's commitments concerning personal-level record keeping in humans. Why do we say 'personal-level'? The experimental literature reviewed above relies primarily on personal-level responses—deliberate responses given in full awareness of explicit instructions from experimenters, sometimes instructions that explicitly ask about subjects' conscious experiences. But, for our purposes, it need not be the case that all of the failings revealed by this literature count as personal-level failings. We proceed to argue that some empirical evidence supports a subpersonal reinterpretation of Olsson's framework. To the extent, however, that the literature surveyed above reveals *sub*personal mnemonic failings, this dampens the prospects of even a subpersonal reading of Olsson's framework. Recall, though, that the compound thesis of the present subsection is conditional in the relevant respect: Olsson's personal-level story concerning the way in which justification is connected to value-qua-stability is implausible; and if any version of that story is plausible, it is a subpersonal one. This thesis is consistent with there being no plausible version of Olsson's story about the way in which justification gives rise to value-qua-stability.

What empirical support might there be for a subpersonal reading of Olsson's proposal?[11] We must concede that, given the determination of the personal by the subpersonal, every personal-level failure is, in some sense, a subpersonal one. Thus, regardless of how we categorize subjects' responses in the various experiments alluded to above—as personally or subpersonally governed responses—all of the negative results discussed cut against a subpersonal version of Olsson's story. It's worth noting, however, some reasons for tempered optimism about a subpersonal fleshing-out of Olsson's approach. Cognitive scientists have documented various ways in which the cognitive system tracks temporal patterns and sources, for example, in the Iowa Gambling Task (as performed by normal subjects—see Bechara et al. 2005) and in tasks that exhibit frequency effects (Jones et al. 2013). Also, the grammatical marking of sources of information can affect the accuracy of memory. In some languages, obligatory differences in, for example, verb inflections mark whether the event being reported was seen by the speaker or, instead, learned about by testimony. Such markings enhance subjects' memories for sentences marked as reports of first-hand observations, presumably because subjects treat first-hand observation as epistemically superior to testimonial acquisition, and first-hand verb-inflection triggers application of this bias (Tosun et al. 2013). Such results suggest that the subject tracks at least some of the "sources" of her beliefs subpersonally.[12]

Now consider the Iowa Gambling Task in which subjects receive rewards and penalties for drawing cards from a variety of decks; some decks have a much more profitable structure than others. For instance, in one deck, 10 out of 11 cards pay out $50 per card, while the eleventh shows a loss of $250, which equals an expected utility of approximately $23 dollars per draw. In another deck, 10 out of 11 cards pay $100 per card, while its eleventh shows a loss of $1250, which equals an expected utility of approximately − $23 per draw). Many subjects eventually achieve a conscious aware-

[11] Some readers might wonder whether it's worth pursuing the matter any further. After all, one could cast Olsson's proposal in more straightforward terms: justification confers stability on a belief and stability is valuable—end of story. In contrast, we're inclined to think that any proposed solution to the swamping problem worth its salt must identify, in a relatively convincing manner, the knowledge- or justification-based source of epistemic value, in this case, the plausible relation by which justification produces value-qua-stability. Olsson would seem to agree, and thus he offers his (personal-level) account of the way in which reliabilist knowledge promotes stability—the account we criticize in the main text. In what follows, we argue that a subpersonal version of Olsson's line of thought offers at least as much promise as (and probably more promise than) his personal-level version of it, and we do so in an attempt to identify a plausible connection between justification and value (that is, to identify something specifically to do with knowledge that confers value on knowledge states, by creating stability). We do not, however, attempt to show that *there can be no* personal-level account of the connection between justification and value-qua-stability other than the one Olsson offers. Perhaps such a further personal-level account will be developed, or has been and we have yet to encounter it. We are open to that possibility. Bear in mind that there are at least two important questions in play. First, we should like to know what value-constituting property a knowledge-state might have. The right answer may well be "stability." Second, one might ask "What does a state's being knowledge have to do with that state's being valuable-qua-stable?" We take this to be a deep question about the *epistemic source* of the value at issue, its source in the properties the co-instantiation of which constitutes a state's being knowledge; we focus on the latter question in connection with Olsson's proposal.

[12] In the experiments of Tosun, subjects receive instructions meant to reduce personal-level attention to credibility-related questions: "To make the study phase more similar to a natural language situation in which participants would not be attempting to remember the sentences or the source of evidence, participants were told that the experiment was about their ability to comprehend sentences and their reading times would be measured." (Tosun et al. 2013, p. 125).

ness of the superiority of winning decks over losing decks, but prior to that point (if it ever comes), subjects have no such awareness, yet they nevertheless systematically modulate their selections in favor of the winning decks, while also showing physiological signs of a sensitivity to threat of loss when beginning to reach for losing decks. Thus, prior to the point of personal-level awareness, a subpersonal pattern-tracking process runs in the absence of any corresponding personal-level process. Moreover, the subpersonal process tracks a pattern of financial payout from sources and is thus analogous to a pattern of "epistemic payout" from sources, a pattern of the sort that Olsson's account requires us to track at the personal level.[13]

We acknowledge the limited scope of the evidence of subpersonal source tracking and evaluation. In contrast, the wealth of evidence against a personal-level story of the sort Olsson has in mind seems damning. We close this subsection, then, with a conditional conclusion: If Olsson's strategy pays off at all, it will do so as a subpersonal-level account of processes that produce instrumental value; justified true beliefs are more valuable than merely true beliefs because the operation of certain forms of subpersonal processing increases the likelihood that a subject will continue to act on a true belief when it's produced by a reliable (and thus justified) process, as compared to beliefs produced by unreliable processes.

3.2 Virtue epistemology

As a personal-level account of the source of epistemic value, virtue epistemology holds apparent promise, for it seems to have the resources to articulate a cogent, personal-level solution to the swamping problem.[14] According to a virtue-based approach, knowledge is true belief the correctness of which is *because of*, or which manifests intellectual virtue on the part of, the agent (Zagzebski 1996; Sosa 2007; Greco 2010; Haddock et al. 2010). In this section, we argue, however, that a subpersonal variation on the virtue epistemologist's proposed solution to the swamping problem fares better than an exclusively personal-level account.

[13] Our point is *not* that two processes, a personal-level process and a subpersonal level one, run in parallel and that the subpersonal process is the more metaphysically or explanatorily fundamental of the two, providing the genuine explanation of epistemic value, in opposition to a personal-level explanation that might also be able to do the job. Rather, there is *no* personal-level process of the sort Olsson describes (that's what we take the empirical evidence to have shown); such a process appears *only* at the subpersonal level, if it appears at all. Thus, our criticism of Olsson's personal-level proposal in no way rests on any claim about causal-explanatory exclusion or the relation between realizers and realized states or between supervening properties and their supervenience base. Note that this is an example of structural mismatch. There is no appropriate record-keeping structure at the personal level, so nothing about the personal-level fact of a (true) belief state's being justified accounts (in Olsson's suggested way) for its production of value (i.e., of stability). In contrast, that connection does stand a reasonable chance of appearing at the subpersonal level (or so we've argued). Bear in mind that we seek a kind of solution to the swamping problem that turns us away from potential conceptual truths about the connection between justification and stability (or any other source of epistemic value). Our criterion of adequacy requires that we understand how it is that in many cases involving actual humans, justification (or warrant) adds value to merely true belief, but without committing ourselves to any necessary truth about such addition.

[14] This point has been conceded in various places by Pritchard (e.g., 2009a, b, 2010), who is a leading critic of robust virtue epistemology.

For ease of exposition, we focus on John Greco's (2010) canonical presentation of the virtue-theoretic response to the swamping problem. According to Greco, knowledge is a cognitive success (i.e., the attaining of a true belief) that is because of cognitive ability. Furthermore, *achievements* are defined more generally as successes that are because of ability. Thus, knowledge is a cognitive achievement, the achievement of a true belief reached through ability. Achievements are valuable for their own sake (in a way mere lucky successes are not); therefore, *knowledge*, qua achievement, is valuable for its own sake.[15]

Let us grant for the sake of argument that knowledge is a cognitive success because of cognitive ability, and thus, that knowledge is always and everywhere a kind of achievement (it is success because of ability). Even on these assumptions, the thesis that knowledge is valuable for its own sake, in a way that mere (unknown) true belief is not, follows only if being reached *through ability or virtue* suffices to make a true belief more valuable than its nonknown counterpart. But why should this be?

At this juncture, Greco takes a nod from Aristotle. In the *Nicomachean Ethics,* Aristotle distinguishes between achieving an end through luck and achieving the end through the exercise of one's abilities (or virtues). The latter, according to Aristotle (as Greco 2010 summarizes):

> is both intrinsically valuable and constitutive of human flourishing … In this discussion Aristotle is clearly concerned with intellectual virtue as well as moral virtue: his position is that the successful exercise of one's intellectual virtues is both intrinsically good and constitutive of human flourishing (2010, pp. 97–98).

The claim that "the successful exercise of one's intellectual virtues is intrinsically good" is put forward as an explanation for why knowledge, conceived of as a kind of successful exercise of intellectual virtue, is valuable in a way that unknown true belief is not. But to say that successful exercise of intellectual virtue is *intrinsically* valuable means just this: that exercising one's intellectual virtues is good for its own sake in virtue of properties that are internal to the successful exercising of intellectual virtue.[16]

Now, this may be where explanation comes to an end; perhaps there's nothing helpful to say except that the successful exercise of virtue is valuable because of whatever of whatever intrinsic properties it has that make it valuable. It is not unreasonable to wish, however, for something more, to hope for elucidation of the relation between the successful exercise of cognitive virtue and epistemic value. In the remainder of this section, we argue that a move to the subpersonal level does shed further light on the matter; it illuminates at least part of the source of the value in question, in a way that an entirely personal-level virtue-theoretic proposal, on its own, does not.

Consider one way a virtue epistemologist might add meat to the personal-level account of epistemic value. Intellectual virtues, as such, must be truth-oriented dispositions that are appropriately *cognitively integrated* (e.g., Pritchard 2010; Greco 2010, p. 156, *passim*) within the agent's cognitive character, a point that is embraced else-

[15] This is a condensed version of the argument found in Greco (2010).

[16] This way of thinking about intrinsic value owes originally to Moore (1903). For a more recent extended discussion, see Rabinowicz and Ronnow-Rasmussen (2000).

where by Greco himself when distinguishing true beliefs reached through virtues from true beliefs reached through reliable but ‘strange and fleeting processes’, the exercise of which issue beliefs that, even when true, fall short of knowledge.[17] Cognitive successes that involve “the successful exercise of intellectual virtue” (i.e., that which the virtue epistemologist tells us is intrinsically valuable) are thus cognitive successes the formation of which is grounded in truth-oriented dispositions that are stable and *integrated*, as opposed to being merely fleeting or disintegrated. The properties of a truth-oriented disposition in virtue of which it is cognitively integrated within the agent’s wider cognitive character are thus properties in virtue of which it is valuable. What are *these* properties?

These appear to be subpersonal properties. Here we consider two ways in which cognitive integration, of the sort adverted to by the virtue epistemologist, appears as a subpersonal phenomenon. The first involves the very nature of cognition. The second involves processes by which individual states or abilities become integrated into an existing cognitive system.

Questions about cognitive integration and cognitive systems have arisen forcefully in the recent debates in the philosophy of mind, particularly in connection with the Extended Mind Hypothesis (EMH) (Clark and Chalmers 1998) and the proposal that groups sometimes constitute cognitive systems (Hutchins 1995; Huebner 2013). A prominent thread in the debate over EMH can be summarized as follows: To the extent that Clark and Chalmers consider personal-level states (in the context of, e.g., their discussion of Otto and his notebook), their arguments for the extended view bog down (Rupert 2004, 2009, 2013). If, as they indicate, they wish to support the extended view of the mind by appeal to its causal-explanatory superiority—one that privileges natural kinds typed coarsely enough to include a significant number of real-world instances with partly external minimal supervenience bases—there must be a successful science of personal-level cognition that individuates cognitive state-types very coarsely. But, there’s very little extant science of this sort; cognitive science tends to produce fairly fine-grained models.

At this juncture, one naturally turns to cognitive science in search of a boundary that distinguishes cognitive from noncognitive causal contributors to the production of intelligent behavior. One such strategy appeals to the line between causal contributors that appear within the relatively persisting, relatively integrated cognitive system (as a whole—that is, the system roughly equivalent to the individual’s entire mind or self) and those that appear beyond the boundary of that integrated system (M. Wilson 2002; °Rupert 2009). This requires, however, some specification of a measure of integration. Rupert appeals to various conditional probabilities of mechanisms’ co-contribution to the production of the subject’s intelligent behavior. The view is not without its problems (Klein 2010; de Brigard 2017), and competing proposals have been made. For example, drawing on the work of Sporns and his colleagues (Sporns et al. 2004), Goldstone and Gureckis appeal to a measure of computational complexity (Goldstone and Gureckis 2009, p. 428; also see Clark 2008, p. 251, n24) to characterize the sort

[17] See here Greco’s (2010, p. 156) diagnosis of Plantinga’s (1993) brain lesion case. For a related discussion of cognitive integration and its connection with virtue epistemology, see Pritchard (2010) and Menary (2012). For discussion of cognitive integration in the context of the extended mind debate, see Menary (2010).

of integration characteristic of an integrated cognitive system. And, Edwin Hutchins proposes that a steep drop-off in the computational gradient ("steep gradients in the density of interaction among [representational] media") marks the boundary of the cognitively relevant unit of analysis (Hutchins 1995, p. 157). For present purposes, we emphasize only that, as it has taken shape, the debate clearly concerns subpersonal-level properties of the cognitive system. For, the states in question do not appear to be of the sort to which folk psychological or rationalizing explanations appeal, and there's no reason to think that subjects have conscious access to the states or properties in question. For example, the sort of computational complexity at issue for Sporns and Goldstone and Gureckis concerns such features as the density and clustering of information-passing channels that connect different components of the cognitive architecture, such as whether those patterns consist in a so-called small-world architecture (dense local clustering with a small number of "long-range" connections) or whether the various computationally specialized subunits are fully connected—every one connected directly to every other. Hutchins gives no indication of thinking that subjects can identify by introspection the fact that the genuine components of their cognitive systems are the ones within the boundary set by the steep gradient in computational processing.[18] Thus, to the extent that progress is being made on issues of cognitive integration, in connection with the extended-mind debate, it is only where contributors "descend" to the level of subpersonal processes.

Now consider a different sort of integration, the way in which newly acquired skills and memories are integrated into the subject's cognitive profile. One especially striking stage of the integration process occurs during sleep. Sleep consolidates skills and memories, and it does so in a way that allows a new motor routine or the content of new experiences to be incorporated into the cognitive system's overall functioning; this is partly a matter of maintaining balance with and facilitating behavior-controlling cooperation with other bodily skills and other parts of one's store of memories. The latter case often goes under the heading of memory consolidation, the process by which memories are cemented (relatively speaking), in contrast to being lost or eliminated (as the records of most of our experiences are). A central and relevant aspect of memory consolidation is described by Dudai, Karni, and Born: "Consolidation is a dynamic, generative, transformative, and lingering process that is posited to balance maintenance of useful experience-dependent internal representations of the world with the need to adapt these representations to the changing world" (2015, p. 21). And, this consolidation process is of particular importance in the case of propositional knowledge: "There is also growing evidence that this sleep-associated redistribution of information is accompanied with an increased semantization of memories and the

[18] Compare Tononi's phi-based theory of consciousness. Although Tononi takes consciousness to consist in a certain sort of informational integration, he does not claim to have arrived at that theory by introspection. A subject might be able to report, de re, on variations in informational integration, simply by reporting on the extent to which a percept seems vivid or is similar to another; but the subject has no conscious access to the fact that what she is reporting on is variation in the highly complex quantitative measure phi (Tononi 2008, p. 220). Phi is the reduction- or supervenience-base of the property of something's being conscious, not the introspectively available content of a personal-level conscious state. Note too that Clark's interest in measures of integration reflects his interest in a story about the realizers or "local material supervenience base" (Clark 2007, p. 186) of personal-level states; that's what's at issue in his disagreement with Rupert (and others).

abstraction of gist information from episodic representations" (*ibid.*, 23), and "the hypothetical process of systems consolidation is most commonly discussed within the context of declarative memory" (*ibid.*, 26). Moreover, the reconsolidation process appears to be directed at, and triggered by the need to, integrate declarative memory into existing bodies of represented facts, of the sort relevant to inference: "Hence, one may hypothesize that, instead of external cues, reactivated pre-existing schemas in neocortical sites direct sleep-dependent consolidation, for example, by favoring the hippocampal reactivation of that memory information that fits the preexisting schema" (*ibid.*, 25), "[P]rior knowledge schemata shape the engagement of the hippocampus in declarative consolidation…" (*ibid.*, 26), and "At the same time, one should not overlook the postulated role of consolidation in balancing stability and change and maintaining adaptive predictive power of representations" (*ibid.* 28).

These may be strange-sounding processes, but their basis is not fleeting; it reflects fundamental operations of the cognitive architecture. Furthermore, the processes in question clearly are not at the personal level; subjects have no conscious access to the fact that copies of recordings of experiences, temporarily stored in hippocampus are being re-encoded in frontal cortex; and that fact is not recognized by folk psychology or adverted to in rationalizing explanations of action. There's no sense in which the *agent* herself shunts those traces from one bit of cortex to another, except the degenerate sense in which the agent does everything that happens at the subpersonal level, such as detecting zero-crossings in early visual processing (Marr 1982). And, notice that it's not simply a matter of cementing memories or practiced routines; it's simultaneously the maintenance of all of the subject's existing cognitive activities and skills; for the incorporation of anything new into that system involves the careful adjustment of relations among existing structures as well as relations to new ones, so as to maintain the integrated functioning of the entire system, including its justification-related functioning, for example, in inference.[19]

In conclusion, if we wish to understand the kind of cognitive integration appeal to which fills the large explanatory lacuna in the virtue-epistemologist's proposed solution to the swamping problem, we do best to look—particularly at the point where Greco's explanation bottoms out—to the subpersonal level. Only there, it seems, do we find the structure and complexity that adds significant explanatory power to the virtue-epistemologists appeals to integration, such structure and complexity as sheds light both on what constitutes the appearance of a single integrated set of cognitive virtues and how such integration is dynamically maintained within a single cognitive system, in response to new information or pieces of evidence. On the virtue-based account, value flows from the exercise of intellectual virtues, which must be understood as part of an integrated psychology, integrated with regard to knowledge structures and to cognitive abilities. Such psychological integration is, however, a subpersonal matter.

[19] For more on memory consolidation, see Rasch and Born (2013) and Squire et al. (2015). Approaching the issue of integration from a slightly different angle, consider the problem of catastrophic interference, which afflicts many neural network models of learning and remembering; this problem arises when the changes in weights involved in the storage of a new pattern "overwrite" weights that encode previously stored patterns or associations. See Ans et al. (2004) and Srivastava et al. (2014) for (clearly subpersonal) attempts to solve this problem in what might be neurologically realistic ways.

4 Straightforwardly subpersonal processing

In this section, we argue directly for subpersonal solutions to the swamping problem, *sans* any detour through extant, personal-level accounts of the value of knowledge. Presently, we develop two arguments each of which rests on a subset of the following background assumptions:

> *Correspondence*: In order for a belief to be true, its propositional content must correspond to reality.
> *Compositionality*: It is a necessary condition on a belief's being a belief that *P* that the belief's components represent, express, or refer to the individuals, relations, properties, etc. constitutive of *P*.
> *Belief Endurance*: In order that a subject's belief be the same belief at two points in time, it must be an attitude toward the same proposition at both of those points; more generally, for any two beliefs B1 and B2 (separated in whatever way) to be type-identical beliefs, B1 and B2 must be attitudes toward the same proposition.[20]
> *Subvenience*: It is a nomologically necessary condition for a belief's continuing to be held that the subject who holds it continue to be in some subpersonal state(s) or other that subvenes (or realizes) the belief in question.
> *Ongoing State*: Although the *acquisition* of a new belief might be an *event with a terminus*, having a belief is itself an *ongoing state* (Vendler 1957).
> *Maintenance*: Having a justified belief is not only a matter of having a belief acquired under appropriate circumstances; it is also a matter of sustaining that belief in an appropriate way.

A few words are in order regarding these assumptions. Firstly, note that *Correspondence, Compositionality*, and *Belief Endurance* are all implied by orthodox thinking about various aspects of the possession conditions for true beliefs (at a time or over time).[21] The remaining three assumptions require elaboration. *Subvenience* articulates a necessary condition for belief retention, namely, that the continued existence of a belief depends on the continued existence of an (that is, some or other) appropriate subvening base.[22]

[20] *Compositionality* and *Belief Endurance* together provide a plausible path to belief-alteration: the representation of a component of *P* might change, which changes the belief in question (it is no longer a belief that *P*). We acknowledge the possibility, however, that the content of a belief might change in some other, more holistic manner. The second argument below, put specifically in terms that presuppose a combinatorial semantics for belief states, might well be recast in a way not so focused on the changes in the referents of "sub-sentential" components, although we make no attempt to work out such an alternative formulation here.

[21] Note, moreover, that our use of 'proposition' is meant to be neutral with respect to the metaphysics of propositions; our assumptions align with what Cappelen and Hawthorne (2009) call 'The Simple View' according to which propositions are taken at least to play certain functional roles characteristically attributed to propositions—viz. as the primary bearer of truth values, the objects of agreement or disagreement, etc.

[22] The subvenience assumption, it should be clear, is stated at a level of generality such that it is applicable to *occurrent beliefs* and *dispositional beliefs* alike. Even if one is not occurrently believing a proposition *P*, one may nonetheless dispositionally believe *P*, provided one is disposed to affirm *P* and has the relevant content stored in memory (which allows for a variety of supervenience bases). If the memory trace is lost, so is the dispositional belief. For further discussion of this distinction, see Schwitzgebel (2015, §2.1). Note

Ongoing state and *Maintenance*, which one of us has defended in detail in previous work, should not be controversial. Nonetheless, some contemporary writers on epistemic value obscure the point. In particular, consider again Zagzebski's analogy with coffee production. The idea in play was that a good-tasting cup of coffee takes on no additional gustatory value simply in virtue of its being the product of a reliable coffee machine.[23] From this point, we are invited to conclude—by parity of reasoning—that it is unclear how a true belief would become additionally epistemically valuable if turns out that the true belief was not only true, but also the product of a reliable belief-forming process.

Consider a different analogy. The project of maintaining a pleasant home is rather unlike the project of making a cup of coffee. The property of being well-maintained—even though it contributes instrumentally to the home's being a pleasant home—is not a property that could be 'swamped' by the value of an already pleasant flat. After all, *if the flat is going to continue to be pleasant, it will have to go on being well-maintained* as the home continues to persist; and accordingly, the property of being well-maintained can continue to confer value to the home indefinitely (Carter et al. 2013, p. 256). What separates beliefs that are candidates for knowledge (i.e., ones which are justified) from mere true beliefs is precisely what separates (by analogy) more generally ongoing states that are positively evaluable from those that are not; the former are sustained through good maintenance that the latter are not.

We propose that oftentimes (even if not always) subpersonal, justification-conferring processes—of the very sort that can help to explain why true beliefs arrived at via cognitively integrated virtues are more valuable than otherwise—underlie the continuing *existence* of a belief.[24] In many cases, for example, a belief (or perhaps better, a proto-belief) does not persist absent certain subpersonal justificatory processes. The belief no longer exists—either by the elimination of its subvening states or by the alteration of its content so as to make it a different belief state—if it is not effectively justified *in an ongoing manner*. In fact, in a wide range of cases, it is highly unlikely that a human subject has—for very long, anyway—a merely true belief. After all, if it is a true, belief-like state *but does not become cemented by subpersonal processing that is also justification-conferring*—such as memory-consolidation—it is oftentimes eliminated, either by a change in content or by the elimination altogether of its subpersonal basis.

Footnote 22 continued
that dispositional beliefs are importantly different from *dispositions to believe.* The content apposite to the former must be at least stored in memory for the dispositional belief to persist. In contrast, a subject may have dispositions to believe (but not dispositional beliefs) contents she has never explicitly represented. For the canonical presentation of this distinction, see Audi (1994).

[23] The thought seems to be that, once one has a good-tasting cup of coffee, it remains good, or to the extent that it does not, the degradation of flavor likely results from, for example, chemical interactions with the surrounding gas molecules, independent of the production process; the goodness of the flavor of the two cups of coffee is, by hypothesis, a function of the appearance of the same chemical profile in them when the process that created them terminates, which screens off, from the differing processes of production, the later chemical interactions that lead to the degradation of flavor.

[24] Note that such justification-conferring subpersonal processes may allow for the subject to be epistemically responsive (e.g., to potential signs of unreliability) even if the subject never becomes consciously aware of any signs of unreliability. These are points that have been echoed in the literature on predictive processing and the Bayesian brain (e.g., Clark 2015).

Argument 1

A1. Premise 1. In many cases, in order that an initially formed belief-like state be maintained long enough for it to become a full-fledged belief state, its subpersonal realizer(s) must be integrated into the cognitive system.

A1. Premise 2. In order that a belief endure over a significant period of time, its subpersonal realizer(s) must continue to survive integration-related subpersonal routines.

A1. Premise 3. The kinds of integration referred to in the two preceding premises contribute to the belief's justificatory status in ways that have no structural parallel at the personal level.

A1 Premise 4. In contrast, a belief-like state that is merely true is typically not integrated into the cognitive system or is not maintained after formation, and thus is likely either never to become a full-fledged belief or to be eliminated in relatively short order.

A1 Premise 5. On the assumption that being true provides some noninstrumental value, a justified true belief has—diachronically—more of it than a nonexistent belief or a belief-like state that is eliminated after a brief existence.

A1 Premise 6. Given A1 Premise 3, the account of this difference in value is distinctively subpersonal.

Therefore, in the cases in question, there is a straightforward—even if not traditionally explored—sense in which justified true beliefs are more noninstrumentally valuable than what would be the relevant merely true beliefs; and the account of this noninstrumental value is distinctively subpersonal.

Let us clarify and elaborate on three points. Firstly, note that scope of the first premise—viz., 'in many cases'. This aligns our argument with the criterion of adequacy defended at the outset. We are not attempting to show that, necessarily, for any subject and any *P*, her knowledge state that *P* has more epistemic value than would her merely true belief that *P*. But, neither do we intend to show merely that it's possible that there be a subject such that some of her knowledge states are more valuable than would be the corresponding merely true belief. Rather, we mean to establish a substantive conclusion intermediate in strength and relevant to the human pursuit of knowledge: that for actual human subjects (and presumably those in nearby possible worlds), a great many of her knowledge states are such that they are more valuable than would be her merely true belief with corresponding content.

Secondly, we should make clear what sort of belief-cementing and belief-maintaining cognitive processes we have in mind. Central to our conception of such processes are those described in the final portion of the preceding section, pertaining to memory consolidation. The role of such processing clearly supports A1P4, for if memory consolidation and the related justification-related integration do not occur, the memory in question is eliminated; if it's not cemented, it simply vanishes. That's the nature of memory consolidation; what is not consolidated is lost (although a nuanced treatment of the point would of course deal in probabilities).

But, to add further depth to the discussion, consider the process of checking for consistency. One might imagine this would occur consciously, via the conscious contemplation of the relation between one's newly acquired belief (or belief-like state) and the rest of one's beliefs. One concentrates on the content of the belief and the contents of one's other beliefs and checks the set for consistency or other, perhaps more robust, coherence-related relations. To the extent that one has justification for one's existing beliefs, consistency or coherence with them provides justification for a newly formed belief (or belief-like state).

There's good reason to think that humans do not implement anything in the vicinity of this personal-level ideal. According to Christopher Cherniak's calculations, the combinatorial explosion entailed by any effort to check explicitly for consistency would sink any such effort (Cherniak 1986). The subpersonal cognitive system instead uses all manner of computing tricks, typically beyond the ken of the conscious mind, to try to maintain consistency, without depending on deliberate, conscious, serial, personal-level calculation. Some such processes occur during slow-wave sleep, as part of a process of reactivating and strengthening representations of facts. As a result of this process, the beliefs in question will not only be justified, but will also be in a position to justify other beliefs, a position that is explained by subpersonal processing. Moreover, this process puts a belief in a position to be further justified by other personal-level states. Your observation of a new dog in the neighborhood might undergo consolidation and integration in a way that creates justificatory power for your resulting belief B. But, it also situates B in a collection of beliefs, including other beliefs about dogs, pets, ownership, etc., such that those other beliefs are more likely to maintain or increase justification for B, when appropriate. This results partly from a declarative memory's integration into existing knowledge-schemas, as happens during the consolidation (and reconsolidation) process described by Dudai et al. (2015). In the cases in question, the only available account of the sources of this justificatory positioning lies at the subpersonal level. Sleep-based memory consolidation involves no personal-level process to which a theorist might appeal.

We do not claim that contemporary cognitive science has yielded a complete and well-confirmed theory of the subconscious processes at issue; but it is highly plausible that any promising heuristic owes its efficacy to subpersonal processing. Consider consistency again. Imagine that a manageable number of randomly selected belief-realizers are activated at various times and, at each time, the active set is subject to a manageable process of consistency checking. How does one select beliefs (or their subvening structures) randomly? It boggles the mind, unless one allows some kind of fast, automatic search process, say, the selective activation of a subset of one's beliefs by the operation of an algorithm that samples from codings of them.[25]

[25] In the field of artificial intelligence, it is relatively common to exploit, in various ways, the strategy of sampling values and inferring from the properties of the sampled set something about the properties of a larger set of data or portion of the world not directly accessible to the agent. (See Russell and Norvig 2011, sectios 14.5 and 15.5.3, which discuss approximate inference in Bayesian networks.) To be clear, the current point is not that the human brain is using, for instance, Markov chain Monte Carlo simulations to maintain epistemic hygiene, only the weaker point that current work in AI at least provides some clue to the sort of strategy that might be used by the brain to maintain epistemic hygiene, given that such hygiene is generally not maintained via conscious reflection on our belief and evidence sets. Note well that such processes, if they occur in humans, are (as memory consolidation processes are) clearly subpersonal. To the extent that the kind of sampling, replaying, and reactivation processes in question appear in humans, they are not available to consciousness and do not appear in folk psychological explanations or in rationalizing explanations that assign normatively governed belief-desire pairs to agents to account for their actions.

Thirdly, we have used the language of being ‘belief-like’ to refer to states that appear in the early stages of belief-formation and that either are or might well become beliefs. We do not insist on this terminology, but we do hold that for many initially acquired or formed states of the belief-like, information-encoding sort, for that state to become a well-functioning belief (a ‘full-fledged belief’ one might say), it must be integrated into one’s cognitive system. Furthermore, this process of integration is justification conferring, because such properties as consistency and broad coherence are justification conferring.[26]

An objector might insist that merely true beliefs possess value indistinguishable from that possessed by justified true beliefs. But as we’ve suggested, there are good empirical grounds to hold in many cases at least, such a merely true belief is likely to crumble and vanish quickly, for what are, from the standpoint of the personal level, inexplicable reasons. To be clear, our concern in this regard is not that a subject will be easily swayed by reasons (good or bad) to give up a merely true belief—because, as it’s sometimes claimed, a merely true belief isn’t, to use the Socratic metaphor, “tied down” at the personal level. Rather, very often—again, even if not always—the belief cannot even be the sort of stable thing that enters into reasons-sensitive relations until it is justified and even then, often, only insofar as it *continues* to be justified at the subpersonal level.[27]

Is the value in question noninstrumental? In some sense, it clearly is. If merely true beliefs are unstable and prone to go out of existence before they even become full-fledged belief-states—compare here with Socrates’ Statues of Daedalus in the *Meno*[28]—integrated into the subject’s psychology, then for any given true belief that *P*, whatever noninstrumental value the merely true belief that *P* has (or would have, in virtue of being true), its justified counterpart has *more* of it, if only because it lasts longer. We are not here committed to the view that there’s an undifferentiated value provided by truth alone, and thus that any long-held, merely true belief—perhaps a trivial one—might provide more value than any justified true belief held for, say, 2 weeks. We make no commitment either way about such cases. Rather, our arguments depend only on the claim that *when the subject and the true proposition are held fixed*, value accrues over time. We hold that, for a given subject, for a given belief that *P*, that subject’s merely true belief that *P* typically has less value than that subject’s knowledge state that *P*, if the latter persists longer than the former. Consider an analogy: if I were to possess Picasso’s *Guernica* for two weeks, I would consider myself lucky; to possess that very same painting for a year would add significantly more value to my aesthetic treasures—even though it may be that there is no length of time such that my owning a mediocre piece of art for that amount of time outweighs the value of owning *Guernica* for two weeks. And, to describe the underlying comparison dramatically, the contrast, at least in some cases, is not between *S*’s holding the mere true belief that *P* and *S*’s holding a justified true belief with the corresponding content *P*; it is between *S*’s

[26] See, for discussion of the value of such broad coherence, Sosa (1997).

[27] This conclusion comports well with Ernest Sosa’s (1991) account of the value of what he calls reflective knowledge, a value that derives from a true belief’s situatedness in a broadly coherent network of other true beliefs.

[28] These statues lacked a certain value, in and of themselves, given their disposition to run away if not tethered down. The value of the statues is realised only in the presence of appropriate tethering.

having no such merely true belief at all and *S*'s holding its justified counterpart. Clearly the latter state of the subject is more noninstrumentally valuable, if there's anything noninstrumentally valuable at all about having the true belief in question; any positive amount is greater than zero. And, less dramatically, in cases in which a merely true belief exists but only for a brief time, its justified, longer-lived counterpart is of greater value. By virtue of its existing as a true belief for a longer period, the latter accrues a greater amount of whatever sort of noninstrumental value truth confers.

Another potential objection holds that the sorts of consolidation processes we have in mind do not confer justification on beliefs, even if such processes enhance internal consistency and coherence in the subject's overall cognitive profile. After all, some such processes serve merely to increase the coherence and maintain the consistency of what to most of us would seem to be a ridiculous set of beliefs (pertaining, say, to the subject's conviction that alien abduction has occurred). Presumably, though, such collections of beliefs (about alien abduction and the like) are thought to be ridiculous because they are false, which pushes them outside both sets of beliefs that we mean to be comparing—merely true beliefs and their counterpart knowledge states. Thus, we can fairly set this concern aside.

We consider now a second argument:

Argument 2

A2. Premise 1. In many cases, subpersonal processes mediate the initial fixation of the content of a mental representation and also sustain the relations that keep its representational value fixed (by getting the mental representation into or keeping it in the content-determining relation to property, kind, or individual represented).[29]

A2. Premise 2. Such relations often involve diagnostic relations among internal representations and as such are, loosely speaking, inferential; schematically, mental representation 'a' tracks As because (1) As reliably exhibit feature B, (2) mental representation 'b' is causally sensitive to the presence of B, and (3) the activation of 'b' tends to cause the activation of 'a'.

A2. Intermediate conclusion 1. Subpersonal processes partly causally determine the identity of proposition believed, and thus the belief itself, by determining some of the elements of the proposition believed (by Premises 1 and 2 and *Compositionality*).

A2. Intermediate conclusion 2. In some cases, subpersonal cognitive processes maintain a belief (or belief-like state), keeping it in existence by grounding the tracking relations that determine the semantic content of components of the representational structure that picks out the proposition the belief in question is an attitude toward (a change in which would eliminate that belief—per *Subvenience* and *Belief Endurance*).

A2. Premise 3. The internal causal relations between subpersonal states, which relations support the relevant tracking capacities, contribute to the belief's (externalist) justificatory status in ways that have no parallel structural at the personal level.

[29] Such diagnostic or mediating relations need not be—in fact, in most cases are not—definitional relations or expressions of necessary and sufficient conditions. On non-defining sustaining mechanisms that help to fix and maintain the content of mental representations, see Fodor (1987, p. 121), Margolis (1998), and Cowie (1999).

> A2 Premise 4. In contrast, a merely true counterpart belief is likely to go out of existence; for if it does not enter into the justification-related relations described—for instance, subpersonal relations of diagnostic inference (or "inference")—then it is not likely to retain its content.

Therefore, for reasons parallel to those given in connection with Argument 1, a justified true belief is, at least in many cases, more noninstrumentally valuable than a merely true belief.

In support of A2, Premise 1, note that many of the ways in which humans skillfully track individuals, property-types, and kinds in the environment rely on the detection of so-called microfeatures, such as subtle differences in gait, scent, or silhouette. Perhaps some such features can be articulated, but often—for example, in face recognition—we can't report on how reidentification occurs. We simply do it. The parents of identical twins, for instance, often can tell the twins' faces apart, even when others can't, but cannot accurately report how they do it.

In support of A2, Premise 2, we note that the relations in question diagnose the presence of the individuals, kinds, or properties represented by the subpersonal units on which the belief content supervenes and that this "diagnosing" relation is generally of epistemic value. Consider a case in which one recognizes the presence of one's pet dog by slight differences in gait or scent, differences that one has difficulty articulating or bringing clearly to consciousness. In such a case, it is because of the association between subpersonal representations that one is plausibly justified in believing that Fido is in the house. Moreover, Bayesian models of cognitive processing predominate at the subpersonal level, which at least *prima facie* involve a justification or confirmation-relation in response to the environment (see Clark 2015 for more on one important class of Bayesian models).

In support of A2, Premise 4, consider at least the possibility that even slight changes in detection routines—changes unreportable by the subject—could alter belief-content, such that some beliefs would go out of existence; this seems especially relevant in cases involving domains in which the targets of the beliefs in question are barely discriminable, such as domains of expert perceptual discrimination. Imagine someone who takes a botany course and who learns, but quickly forgets, what the experts say about the different properties of tree leaves of very similar looking species. How might she maintain beliefs about the properties of various species? Possibly, by being able to track the different leaves and having the belief that *that* one can be eaten, in a pinch, but *that* one should not be consumed under any circumstances. If, however, her ability to discriminate between microfeatural differences of the different kinds of leaves has suffered from degradation or interference, she no longer has the beliefs in question. If, in contrast, the subpersonal relations in question are maintained, so too are the relevant beliefs.

Of course, accessibilist internalists (Chisholm 1988; BonJour 1985) will likely express doubts about the appeal to subpersonal processes. They are almost sure to claim that subpersonal processes have nothing to do with justification, that consciousness must have direct access to anything that counts as justification and that subjects do not have conscious access to the states, relations, and processes in question.

Consider, however, the following possibility. Imagine that various subjects hold various beliefs that they take to be self-justifying or justified a priori (though perhaps fallibly so)—most importantly, justified in the absence of personal-level justifying relations. Now imagine that a pattern emerges: we secure evidence that a significant subset of these beliefs are false and that there is a principled distinction between the subpersonal processes that lead to the formation and maintenance of the true ones, on the one hand, and the subpersonal processes that lead to the formation and maintenance of the false ones, on the other hand. Focused on the personal level alone, the accessibility internalist can appeal to no structure or process that would explain the difference between the two kinds of cases; subjects report equal levels of certainty attached to both kinds of beliefs, and they report not having based them in any way on inferences. In this case, the internalist should accept that a mismatch argument establishes a role for the subpersonal in the theory of justification. We maintain that internalists should be similarly moved by mismatches of the sort we have discussed. In many cases, the processes that validate the justificatory status of states accessible to consciousness are subpersonal and lack any personal-level analogue.

Consider a final objection. One might worry that our discussion of subpersonal phenomena—in particular, reasons for thinking that the sources of epistemic value reveal themselves only (or at least largely) subpersonally—ignores broader conceptual issues. What is it, one might wonder, about knowledge itself, that makes it valuable, setting aside the way it appears or is maintained in a particular kind of physical system. In fact, one might wonder whether all that we have said about the empirical cases can be ignored. To the extent that we have addressed the swamping problem, it is by saying something of the following sort: "justification adds value because it enhances stability (or personal-level integration, or increases the amount of time one holds a given true belief)." And, *that* could have been said without any detour through the empirical literature or the exploration of subpersonal processing.

Here's the rub, though. It's not a conceptual truth that justified true beliefs are more stable or less likely to go out of existence or more integrated than mere true beliefs, and thus it does not suffice to solve the swamping problem merely to point to these (possibly) personal-level traits. Sure enough, *if* justification is correlated with some valuable characteristic F—whether it's stability or cognitive integration or longer-lasting truth—then justification adds value to true beliefs. But, we would like to know whether, *in fact*, a belief's being justified is correlated with the presence of F. That might hold for some creatures, depending on the way they're built, but not for others. We would especially like to know whether humans are built in such a way that our justified true beliefs are more stable, etc. than our merely true beliefs (and perhaps thereby whether it's possible for any creatures to be built in that way). We would like to know that knowledge is in fact often valuable *for us* (and more valuable than mere true belief), and one path to such a result is to see that the processes by which justification appears in our case are correlated with stability, cognitive integration, or longer-lasting truth.

Before closing, it will be helpful to register two summary points about the conclusions we have reached and to situate them in the context of the initial value of knowledge debate with which we began. Firstly, and with reference to the taxonomy introduced in the Introduction—including *validationism, fatalism* and *revi-*

sionism—we take ourselves to have explored two importantly distinct strands of validationism, both of which involve novel recourse to the subpersonal level, and which have been hitherto ignored. The first strand of validationist response showed how two leading attempts to defend validationist strategies—developed by Olsson and Greco—would do well to incorporate subpersonal theoretical components; doing so increases the chances that such approaches can offer viable and satisfying accounts of the value of knowledge *on their own favored terms*. If this is right, then we've shown an important respect in which the role of subpersonal processes has been overlooked and has importance in epistemic axiology.

The second strand of validationist response, offered in Sect. 4, was comparatively more ambitious. According to this strategy of response, the pretheoretical insight that the value of knowledge exceeds that of mere true opinion can in principle be vindicated *exclusively* at the subpersonal level of description. To be clear, we maintain that our case for the philosophical import of the subpersonal in epistemic axiology does not actually require this second strand of validationist response; the first strand would suffice. We have, however, attempted to show how even this stronger strand of validationist strategy has much to recommend it. In doing so, we offered two connected arguments, each of which presented a novel way in which we envisage a subpersonal response to the swamping problem being developed.

Finally, we admit, even emphasize, our limited understanding of the nature of the subpersonal processes in question. But, we take ourselves to have provided sufficient reason to be enthusiastic about the present approach as a way to make progress on questions about epistemic value, generally speaking, and the swamping problem, in particular.[30]

References

Alfano, M. (2013). *Character as moral fiction*. Cambridge University Press.

Ans, B., Rousset, S., French, R. M., & Musca, S (2004). Self-refreshing memory in artificial neural networks: learning temporal sequences without catastrophic forgetting. *Connection Science*, *16*(2), 71–99.

Audi, R. (1994). Dispositional beliefs and dispositions to believe. *Nous, 28*(4), 419–434.

Baehr, J. (2009). Is there a value problem? In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value*. Oxford: Oxford University Press.

Barbar, S. J., Rajaram, S., & Marsh, E. J. (2008). Fact learning: how information accuracy, delay, and repeated testing change retention and retrieval experience. *Memory, 16*(8), 934–946.

Barnier, A. J., Sutton, J., Harris, C. B., & Wilson, R. A. (2008). A conceptual and empirical framework for the social distribution of cognition: The case of memory. *Cognitive Systems Research, 9*(1), 33–51.

[30] The authors are grateful to two anonymous referees at *Synthese* as well as to the editors.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa gambling task and the somatic marker hypothesis: Some questions and answers. *Trends in Cognitive Sciences, 9*(4), 159–164.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General, 121*(4), 446–458.

BonJour, L. (1985). *The structure of empirical knowledge*. Cambridge: Cambridge University Press.

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology, 62*(3), 311–318.

Brown, A. S., & Marsh, E. J. (2010). Digging into Déjà Vu: Recent research on possible mechanisms. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 53, pp. 33–62). Burlington: Academic Press.

Cappelen, H., & Hawthorne, J. (2009). *Relativism and monadic truth*. Oxford: Oxford University Press.

Carter, J. A., Jarvis, B., & Rubin, K. (2013). Knowledge: Value on the cheap. *Australasian Journal of Philosophy, 91*(2), 249–263.

Chambers, K. L., & Zaragoza, M. S. (2001). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory and Cognition, 29*(8), 1120–1129.

Cherniak, C. (1986). *Minimal rationality*. Cambridge, MA: MIT Press.

Chisholm, R. M. (1988). The indispensability of internal justification. *Synthese, 74*(3), 285–296.

Clark, A. (2007). Curing cognitive hiccups: A defense of the extended mind. *Journal of Philosophy, 104*(4), 163–192.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Cowie, F. (1999). *What's within: Nativism reconsidered*. Oxford: Oxford University Press.

Davies, M. (2000a). Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society, 1,* 87–105.

Davies, M. (2000b). Persons and their underpinnings. *Philosophical Explorations, 3,* 43–62.

David, M. (2001). Truth as the epistemic goal. In M. Steup (Ed.), *Knowledge, truth, and duty*, (pp. 151–169). New York: Oxford University Press.

de Brigard, F. (2017). Cognitive systems and the changing brain. *Philosophical Explorations, 20*(2), 224–241.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review, 14*(2), 238–257.

Dennett, D. C. (1969). *Content and consciousness*. New York: Routledge.

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.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

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.

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron, 88,* 20–32.

Echterhoff, G., Hirst, W., & Hussy, W. (2005). How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Memory and Cognition, 33*(5), 770–782.

Fazio, L. K., Brashier, N. M., Keith Payne, B., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General, 144*(5), 993–1000.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. *Evans and Frankish, 2009,* 89–107.

Gendler, T. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634–663.

Geraci, L., & Franklin, N. (2004). The influence of linguistic labels on source-monitoring decisions. *Memory, 12*(5), 571–585.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), *Justification and Knowledge* (pp. 1–25). Boston: D. Reidel.

Goldman, A., & Olsson, E. J. (2009). Reliabilism and the value of knowledge. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value*. Oxford: Oxford University Press.

Goldstone, R. L., & Gureckis, T. M. (2009). Collective behavior. *Topics in Cognitive Science, 1,* 412–438.

Greco, J. (2010). *Achieving knowledge: A virtue-theoretic account of epistemic normativity*. Cambridge: Cambridge University Press.

Haddock, A., Millar, A., & Pritchard, D. (2009). *Epistemic value*. Oxford: Oxford University Press.

Haddock, A., Millar, A., & Pritchard, D. (2010). *The nature and value of knowledge: Three investigations*. Oxford: Oxford University Press.

Harman, G. (2000). The nonexistence of character traits. *Proceedings of the Aristotelian Society*, *100*(2), 223–226.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16,* 107–112.

Haybron, D. M. (2007). Do we know how happy we are? On some limits of affective introspection and recall. *Noûs, 41*(3), 394–428.

Heck, R. G. (2000). Nonconceptual content and the'space of reasons". *The Philosophical Review, 109*(4), 483–523.

Henkel, L. A., & Mattson, M. E. (2011). Reading Is believing: The truth effect and source credibility. *Consciousness and Cognition, 20,* 1705–1721.

Huebner, B. (2013). *Macrocognition: A theory of distributed minds and collective intentionality*. Oxford: Oxford University Press.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Jacoby, L. L., Kelley, C., Brown, J., & Jasechko, J. (1989a). "Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56*(3), 326–338.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989b). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General, 118*(2), 115–125.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28.

Jones, W. E. (1997). Why do we value knowledge? *American Philosophical Quarterly*, *34*(4), 423–439.

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review, 120*(3), 628.

Klein, C. (2010). Critical notice: Cognitive systems and the extended mind by Robert Rupert. *The Journal of Mind and Behavior, 31*(3&4), 253–264.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge: Cambridge University Press.

Kvanvig, J. (2008). Pointless truth. *Midwest Studies in Philosophy, 32*(1), 199–212.

Kvanvig, J. (2010). 'The swamping problem redux: Pith and Gist'. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social epistemology* (pp. 89–112). Oxford: Oxford University Press.

Lau, H. C., Rogers R. D., & Passingham, R. E. (2007). Manipulating the experienced onset of intention after action execution. *Journal of Cognitive Neuroscience*, *19*(1), 81–90.

Loftus, E. (1979). *Eyewitness testimony*. Cambridge: Harvard University Press.

Lynch, M. P. (2009). Truth, value and epistemic expressivism. *Philosophy and Phenomenological Research*, *79*(1), 76–97.

Lyons, J. C. (2019). Algorithm and parameters: Solving the generality problem for reliabilism. *Philosophical Review*, *128*(4), 463–509

Margolis, E. (1998). How to acquire a concept. *Mind and Language, 13*(3), 347–369.

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.

Marsh, E. J., Eslick, A. N., & Fazio, L. K. (2008). False memories. In H. L. Roediger III (Ed.), *Cognitive psychology of memory. Vol. [2] of learning and memory: A comprehensive reference, 4 vols. (J. Byrne Editor)* (pp. 221–238). Oxford: Elsevier.

Mather, M., Shafir, E., & Johnson, M. K. (2000). Misrememberance of options past: Source monitoring and choice. *Psychological Science, 11*(2), 132–138.

McDowell, J. (1994a). The content of perceptual experience. *The Philosophical Quarterly, 44*(175), 190–205.

McDowell, J. (1994b). *Mind and world*. Cambridge, MA: Harvard University Press.

Meade, M. L., & Roediger, H. L., III. (2002). Explorations in the social contagion of memory. *Memory and Cognition, 30*(7), 995–1009.

Menary, R. (2010). The extended mind and cognitive integration. In R. Menary (Ed.), *The extended mind*. Cambridge: MIT Press.

Menary, R. (2012). Cognitive practices and cognitive character. *Philosophical Explorations, 15*(2), 147–164.

Moore, G. E. (1903). *Principia ethica*. Mineola: Dover Publications.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259.

Olsson, E. J. (2007). Reliabilism, stability and the value of knowledge. *American Philosophical Quarterly, 44*(4), 343–355.

Olsson, E. J. (2011). The value of knowledge. *Philosophy Compass*, *6*(12), 874–883.

Plantinga, A. (1993). *Warrant and proper function*. New York: Oxford University Press.

Pritchard, D. (2007). Recent work on epistemic value. *American Philosophical Quarterly, 44,* 85–110.

Pritchard, D. (2009a). Knowledge, understanding and epistemic value. *Royal Institute of Philosophy Supplement, 64,* 19–43.

Pritchard, D. (2009b). The value of knowledge. *The Harvard Review of Philosophy, 16*(1), 2–19.

Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese, 175*(1), 133–151.

Pritchard, D. (2011). What is the swamping problem. In A. Reisner & A. Steglich-Petersen (Eds.), *Reasons for belief*. Cambridge: Cambridge University Press.

Pritchard, D. (2013). *What is this thing called knowledge?*. Abingdon: Routledge.

Rabinowicz, W., & Ronnow-Rasmussen, T. (2000). II-A distinction in value: Intrinsic and for its own sake. *Proceedings of the Aristotelian Society, 100*(1), 33–51.

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiological Reviews, 93*(2), 681–766.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8,* 338–342.

Ridge, M. (2013). Getting lost on the road to Larissa. *Noûs, 47,* 181–201.

Riggs, W. (2008). The value turn in epistemology. In V. Hendricks (Ed.), *New waves in epistemology* (pp. 300–323). London: Palgrave Macmillan.

Riggs, W. D. (2009). Understanding, knowledge, and the meno requirement. In A. Haddock, A. Millar & D. Pritchard (Eds.), *Epistemic value*, (pp. 331–38). Oxford: Oxford University Press.

Roediger, L., & Gallo, A. (2005). Associative memory illusions. In R. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory* (pp. 309–326). Oxford: Oxford University Press.

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy, 101,* 389–428.

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

Rupert, R. D. (2013). Memory, natural kinds, and cognitive extension; or, Martians don't remember, and cognitive science is not about cognition. *Review of Philosophy and Psychology, 4,* 25–47.

Russell, S., & Norvig, P. (2011). *Artificial intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ: Pearson.

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review, 117,* 245–273.

Schwitzgebel, E. (2015). Belief. In Edward N. Zalta (Eds.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/sum2015/entries/belief/.

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, & T. Thornton (Eds.), *Oxford handbook of philosophy and psychiatry* (pp. 1063–1082). Oxford: Oxford University Press.

Sosa, E. (1991). *Knowledge in perspective: Selected essays in epistemology*. Cambridge: Cambridge University Press.

Sosa, E. (1997). Reflective knowledge in the best circles. *The Journal of Philosophy, 94*(8), 410–430.

Sosa, E. (2000). For the love of truth? In L. Zagzebski & A. Fairweather (Eds.), *Virtue epistemology: Essays on epistemic virtue and responsibility*. (pp. 49–62). Oxford: Oxford University Press.

Sosa, E. (2007). *A virtue epistemology: Apt belief and reflective knowledge* (Vol. 1). Oxford: Oxford University Press.

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences, 8*(9), 418–425.

Squire, L. R., Genzel, L., Wixted, J. T., & Morris, R. G. (2015). Memory consolidation. *Cold Spring Harbor Perspectives in Biology*. https://doi.org/10.1101/cshperspect.a021766.

Srivastava, V., Sampath, S., & Parker, D. J. (2014). Overcoming catastrophic interference in connectionist networks using gram-schmidt orthogonalization. *PLoS ONE, 9*(9), 5. https://doi.org/10.1371/journal.pone.0105619.

Sylvan, K. (2017). Veritism unswamped. *Mind, 127*(506), 381–435.

Swinburne, R (1999). *Providence and the problem of evil*. Oxford University Press UK.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin, 215,* 216–242.

Tosun, S., Vaid, J., & Geraci, L. (2013). Does obligatory linguistic marking of source of evidence affect source memory? A Turkish/English investigation. *Journal of Memory and Language, 69,* 121–134.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. Oxford: Academic Press.

Vendler, Z. (1957). Verbs and times. *The Philosophical Review, 62,* 143–160.

Wegner, D. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Belknap Press of Harvard University Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9,* 625–636.

Zagzebski, L. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge: Cambridge University Press.

Zagzebski, L. (2003). The search for the source of epistemic good. *Metaphilosophy, 34*(1–2), 12–28. https://doi.org/10.1111/1467-9973.00257.